GENOMICS-ENABLED CROP GENETICS

EDITED BY: Yin Li, Chuang Ma, Ray Ming and Wenqin Wang

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-893-9
DOI 10.3389/978-2-88966-893-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

GENOMICS-ENABLED CROP GENETICS

Topic Editors:
Yin Li, Huazhong University of Science and Technology, China
Chuang Ma, Northwest A and F University, China
Ray Ming, University of Illinois at Urbana-Champaign, United States
Wenqin Wang, Shanghai Normal University, China

Citation: Li, Y., Ma, C., Ming, R., Wang, W., eds. (2021). Genomics-Enabled Crop Genetics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-893-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 07 May 2021
doi: 10.3389/fgene.2021.687160

Editorial: Genomics-Enabled Crop Genetics

Yin Li[1]*, *Wenqin Wang*[2], *Chuang Ma*[3,4] *and Ray Ming*[5]

[1] *The Genetic Engineering International Cooperation Base of Chinese Ministry of Science and Technology, The Key Laboratory of Molecular Biophysics of Chinese Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China,* [2] *College of Life Sciences, Shanghai Normal University, Shanghai, China,* [3] *State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Xianyang, China,* [4] *Key Laboratory of Biology and Genetics Improvement of Maize in Arid Area of Northwest Region, Ministry of Agriculture, Northwest A&F University, Xianyang, China,* [5] *Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, United States*

Keywords: crops, genomics, transcriptomics, population genetics, germplasm characterization, gene functional studies

Editorial on the Research Topic

Genomics-Enabled Crop Genetics

OPEN ACCESS

Edited and reviewed by:
Lin-Feng Li,
Fudan University, China

****Correspondence:***
Yin Li
yinli2021@hust.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *29 March 2021*
Accepted: *15 April 2021*
Published: *07 May 2021*

Citation:
Li Y, Wang W, Ma C and Ming R (2021) Editorial: Genomics-Enabled Crop Genetics. Front. Genet. 12:687160. doi: 10.3389/fgene.2021.687160

INTRODUCTION

In the genomics era, omics-based technologies have unprecedentedly promoted progress in plant biology, from plant growth and development, plant physiology to molecular genetic studies, and system and synthetic biology. While proteomics and metabolomics are becoming prevalent, genomics and transcriptomics are the most popular and widely used platforms for crop studies due to their rapidly decreased costs, improved sequencing quality, a broad spectrum of applications, and well-established bioinformatic tools. Genetic and functional genomic studies in crops, especially those in non-model crops, have been lagged far behind compared to those in model plant species for a couple of reasons. First, some crops can have a large, complex and polyploidy genome, such as wheat (*Triticum aestivum*) (International Wheat Genome Sequencing Consortium, 2018). Second, while a group of closely related crop species is often comparatively studied or used in breeding programs, they could have distinct genomes and/or ploidy levels, representing further technical challenges for molecular studies. For example, the peanuts include the cultivated peanut (*Arachis hypogaes*, AABB genome), the wild tetraploid peanut (*Arachis monticola*, AABB genome) and two wild diploid peanuts, *Arachis duranensis* (AA genome) and *Arachis ipaensis* (BB genome) (Bertioli et al., 2016, 2019; Chen et al., 2016, 2019; Lu et al., 2018; Yin et al., 2018, 2019; Zhuang et al., 2019). Another example is the cultivated bananas, which are interspecific or intraspecific hybrids between wild diploid *Musa acuminata* (AA genome) and *Musa balbisiana* (BB genome). They have various genotypes, including diploid (AA, BB, and AB), triploid (AAA, AAB, and ABB) and tetraploid (AAAB, AABB, ABBB) variants (D'Hont et al., 2012; Davey et al., 2013; Martin et al., 2016; Wang et al., 2019). Third, for many crops the genomic resources supporting functional studies and molecular breeding are not often available, including high-quality reference genome assemblies, high-density genetic maps, and genomics-characterized populations. Finally, in some crops (such as sorghum), genetic transformation is still challenging, and mutant resources are not well-established.

When synergistically integrated with other omics approaches, genomic technologies can be compelling for crop genetics, representing a technological basis to help mitigate or circumvent the challenges mentioned above in crop studies. The papers included in this Research Topic, Genomics-Enabled Crop Genetics, illustrate this concept. The various studies collected in this Research Topic can be summarized into three major aspects: (1) the theme of "**Genomic technologies promote germplasm characterization**" includes contributions regarding molecular identification, characterization of crop species and accessions with genomics-based methods. (2) The subject of "**Genomic technologies enhance crop population genetics**" showcases the examples of population genetic studies facilitated by the genomic approaches. (3) The topic of "**Genomic technologies enable functional mining of genomic components in crops,**" on the other hand, presents the applications in multiple genomic and transcriptomic databases. These resources are comprehensively integrated to generate functional insights into the genomic components, e.g., genes, miRNAs and *cis*-regulatory elements.

This Research Topic includes thirteen original research articles, one hypothesis and theory paper, one opinion paper and one review article, covering the following three aspects.

GENOMIC TECHNOLOGIES PROMOTE GERMPLASM CHARACTERIZATION

Markers of simple sequence repeats (SSR) or chloroplast DNA are often used to study the phylogenetic relationship between accessions or species within a crop genus. Taking the advantages of RNA-seq that provides sequence information about functional genes in a cost-effective and high-throughput way, Karcı et al. performed transcriptome sequencing of pistachio (*Pistacia vera*), developed 233 genic SSR markers (gSSR) and studied the phylogenetic relationship using 55 gSSR markers from nine *Pistacia* species. This study exemplifies RNA-seq as a tool to contribute to the taxonomy of crop species and their relatives. Qiu et al. assembled the five fescue taxa's chloroplast genomes, including three subspecies of *Festuca rubra*, one *Festuca brevipila*, and one *Festuca ovina*, providing resources to screen fescue germplasm accessions and to refine species identification. With the plastid genome information, Qiu et al. reconstructed the phylogenetic relationship of the *Festuca-Lolium* complex. Synthetic or artificial polyploid hybrid materials within the *Triticum* genus or *Triticum* and its relative species represent essential wheat genetic improvement resources. Cytogenetic techniques can help provide insights into crop genomics, guiding further investigations on certain genomic issues. For example, to better characterize the tetraploid wheat-*Aegilops ventricosa* amphiploid materials, Zhang et al. observed the chromosomal behavior of the progeny plants ($AABBD^VD^VN^VN^V$) derived from crosses between *T. turgidum* (AABB) and *Aegilops ventricosa* ($D^VD^VN^VN^V$) using multicolor Fluorescence *in situ* hybridization (mc-FISH), providing insights into the genome stability of allopolyploidization in the wheat group.

GENOMIC TECHNOLOGIES ENHANCE CROP POPULATION GENETICS

Genomic-based technologies have enhanced the traditional linkage mapping of quantitative trait loci (QTL) and enabled genome-wide association study (GWAS) by developing hundreds of thousands of markers (single nucleotide polymorphism, SNP, e.g., in most applications). In understudied crops, it is cost-effective to develop abundant SNP markers for QTL mapping by using reduced representation sequencing techniques, such as restriction-site associated DNA sequencing (RAD-seq) (Miller et al., 2007), genotyping-by-sequencing (GBS) (Elshire et al., 2011), and specific length amplified fragment sequencing (SLAF-seq) (Zhang et al., 2013). Wei et al. developed an interspecific F_2 population containing 121 individuals, constructed a genetic map of eggplant (*Solanum melongena*) with 2,122 SNP markers and identified 19 QTLs for several morphological traits. This work lays a foundation for the fine mapping of QTLs and marker-assisted selection in eggplant breeding. In another work, Peng et al. used several SNP-identification methods (target enrichment sequencing, TES, RNA-seq and the 48K Axiom *Arachis*2 SNP array) to identify the genomic region and candidate genes controlling nodulation in cultivated peanut (*A. hypogaea* L.). They demonstrate that TES generated the highest number of SNPs, followed by RNA-seq and the SNP array with GBS being the least effective. In this work, TES and the SNP array have comparable costs per SNP per sample, while RNA-seq was the most expensive technique for SNP identification. To discover candidate genes associated with ear morphology in breeding populations, Li et al. identified SNPs for 208 maize inbred lines from two heterosis groups, Shaan A and Shaan B. The further GBS, combined GWAS and selective sweeps identified four genes associated with ear length and fruit length. Genomic technologies not only enhance QTL mapping, but also help in identifying expression QTL (eQTL). Barbey et al. identified SNPs for octoploid strawberry populations using the Affymetrix IStraw 35 Axiom SNP array and mapped 268 eQTLs for 224 genes expressed in the mature receptacle. Many of the eQTLs are known to affect fruit traits that were either described experimentally or validated via transgenic approaches.

GENOMIC TECHNOLOGIES ENABLE FUNCTIONAL MINING OF GENOMIC COMPONENTS IN CROPS

Integration of multiple genomic resources, including but not limited to genetic variation by whole genome resequencing, gene expression by RNA-seq, miRNA expression by small RNA-seq and miRNA targets' cleavage information by degradome sequencing, can significantly enhance our understanding of transcriptional and post-transcriptional regulation in crops. Glazinska et al. created an expression database for yellow lupine (*Lupinus luteus* L.), namely LuLuDB, by combining RNA-seq analysis of small RNA, transcriptome, and degradome libraries, providing analysis-ready information of the NGS data. They further demonstrated the usefulness of the LuLuDB by a showcase of a genome-wide analysis of the *Dicer Like* (*DCL*) gene family and a *miR486-DCL2* analysis. In maize research, Xu et al. integrated 195 small RNA sequencing libraries and 19 degradome libraries. Together with the identification of phasi-RNA and GWAS results, they found many tissue-specific miRNAs and depicted evolutionary implications of small RNAs. Zhao et al. combined the heat-responsive transcriptomes of wheat and the genome-wide identified heat shock elements (HSEs) and show that a particular variant of non-canonical HSE is associated with a larger heat stress response and that the heat stress-responsive genes containing different HSEs are functionally diverged.

In addition to providing large-scale gene functional implications, genomic datasets can highlight a gene of interest within a particular gene family. For example, in a genome-wide analysis of wheat heat shock protein 90 (*TaHSP90*), Lu et al. took advantage of PacBio Iso-seq data and identified 126 isoforms derived from the *TaHSP90* genes. The highly expressed *TaHSP90-AA* genes showed a large magnitude of response to heat stress with differential alternative splicing patterns observed between the three *TaHSP90* homologous copies, extending our understanding of the functional divergence of the HSP family. Many pan-genome studies have revealed extensive genetic variations between accessions within a crop species, including copy number and structural variations. With the three high-quality phased diploid genomes of grapevine cultivars, Cabernet Sauvignon (CS), Carménère (CR), and Chardonnay (CH), Smit et al. compared the terpene synthase (*VviTPS*) gene family between CS, CR, CH, and an Illumina-based reference genome PN40024 (Jaillon et al., 2007; Chin et al., 2016; Minio et al., 2017, 2019; Roach et al., 2018). The in-depth genome-wide comparison of *VviTPS* family identified duplicated gene copies, predicted functions of *VviTPS* by combining sequence homology and established knowledge of more than 40 biochemically characterized *VviTPS* genes (Smit et al.) Zhang et al. summarized published genome-wide analyses of gene families in the cultivated and wild peanut (*Arachis*) genomes (Bertioli et al., 2016, 2019; Chen et al., 2016, 2019; Lu et al., 2018; Yin et al., 2018, 2019; Zhuang et al., 2019). Zhang et al. show that the hidden Markov Model (HMM)-based search of a gene family is rapid and accurate and provides helpful suggestions regarding aspects of gene family analysis.

The abundant genomic resources allow for investigation on the genomic components other than protein-coding genes and non-coding RNAs, such as untranslated regions (UTR) and *cis*-regulatory elements. Tu and Li (2020) developed an RNA-seq analysis method to profile alternative $3^{'}$ untranslated regions ($3^{'}$UTRs), priUTR suitable for crops like sorghum and maize. Profiling of the genes with alternative $3^{'}$UTRs in *Sorghum bicolor* reveals a link between alternative $3^{'}$UTR and RNA N^6-methyladenosine (m^6A) modification, which had also been implicated in a previous maize RNA m^6A profiling experiment (Luo et al., 2020). These papers provide bioinformatic evidence on the relationship between RNA m^6A modification and alternative $3^{'}$UTRs/ polyadenylation. In 2021, a major breakthrough has been made to the link of m^6A and alternative polyadenylation that the longer isoform of Cleavage and Polyadenylation Specificity Factor 30 (CPSF30-L) is the key protein to mediate m^6A regulation of polyadenylation in *Arabidopsis* (Hou et al., 2021; Song et al., 2021). In this Research Topic, Galli et al. reviewed the state-of-art of our knowledge in regulatory regions and their mechanisms in controlling gene expression. Galli et al. further summarized the cutting-edge NGS technologies for detecting accessible chromatin regions (ACRs) and DNA-binding motifs of transcription factors (TFs). Particularly, the pros and cons of several methods for mapping TFs' DNA binding motifs [i.e., Chromatin immunoprecipitation sequencing (ChIP-seq), DNA Affinity Purification sequencing (DAP-seq) and cleavage under targets and release using nuclease (CUT&RUN)] have been discussed and highlighted. DAP-seq is considered a cost-efficient high-throughput method for crop regulome study. In addition, the reference genomes of closely related crops represent important resources for genome evolution studies. Yu et al. selected four pairs of genomes from the four core eudicot plant families, performed genome-wide synteny block comparison and discovered that excision of genes is much more prevalent than pseudogenization during genome fractionation.

CONCLUDING REMARKS

The collection of sixteen papers in this Research Topic reflects the broad spectrum of current research directions in genomics-enabled crop genetics. The current Research Topic also exemplifies that genomic technologies and resources can be applied to a wide range of crop species, from cereal crops, such as wheat, maize and sorghum, to horticultural crops, such as eggplant, pistachio, yellow lupine, fine fescue, strawberry, and peanut. As the contributions to the Research Topic "Genomics-Enabled Crop Genetics" exemplarily shows, a combination of multiple genomic technologies and/or resources can form powerful and comprehensive tools for different aspects of crop genetic studies, suitable for different crop species with distinct applications and emphases. As more genomic resources and techniques are being developed for a variety of crop species, the output will accelerate crop genetic research and, ultimately, promote crop genetic improvement.

AUTHOR CONTRIBUTIONS

YL, CM, WW, and RM co-wrote this editorial based on this Research Topic's contributions. All authors contributed to the article and approved the submitted version.

FUNDING

YL was funded by the Fundamental Research Funds for the Central Universities (HUST 2021XXJS070). The project was supported by National Natural Science Foundation of China Grant 32072008 (to WW).

ACKNOWLEDGMENTS

We thank all authors who submitted their work for this Research Topic, the support of professional editorial staff at Frontiers, and the invaluable time and efforts of reviewers in manuscript evaluation.

REFERENCES

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet*. 48, 438–446. doi: 10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., et al. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet*. 51, 877–884. doi: 10.1038/s41588-019-0405-z

Chen, X., Li, H., Pandey, M. K., Yang, Q., Wang, X., Garg, V., et al. (2016). Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6785–6790. doi: 10.1073/pnas.1600899113

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., et al. (2019). Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant.* 12, 920–934. doi: 10.1016/j.molp.2019.03.005

Chin, C., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi: 10.1038/nmeth.4035

Davey, M. W., Gudimella, R., Harikrishna, J. A., Sin, L. W., Khalid, N., Keulemans, J., et al. (2013). A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter-and intra-specific *Musa* hybrids. *BMC Genom*. 14:683. doi: 10.1186/1471-2164-14-683

D'Hont, A., Denoeud, F., Aury, J. M., Baurens, F. C., Careel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488, 213–217. doi: 10.1038/nature11241

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19397. doi: 10.1371/journal.pone.0019379

Hou, Y., Sun, J., Wu, B., Gao, Y., Nie, H., Nie, Z., et al. (2021). CPSF30-L-mediated recognition of mRNA m^6A modification controls alternative polyadenylation of nitrate signaling-related gene transcripts in *Arabidopsis*. *Mol. Plant* 14, 688–699. doi: 10.1016/j.molp.2021.01.013

International Wheat Genome Sequencing Consortium. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 361, eaar7191.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature0 6148

Lu, Q., Li, H., Hong, Y., Zhang, G., Wen, S., Li, X., et al. (2018). Genome sequencing and analysis of the peanut B-genome progenitor (*Arachis ipaensis*). *Front. Plant. Sci*. 9:604. doi: 10.3389/fpls.2018.00604

Luo, J., Wang, Y., Wang, M., Zhang, L., Peng, H., Zhou, Y., et al. (2020). Natural variation in RNA m^6A methylation and its relationship with translational status. *Plant Physiol*. 182, 332–334. doi: 10.1104/pp.19.00987

Martin, G., Baurens, F. C., Droc, G., Rouard, M., Cenci, A., Kilian, A., et al. (2016). Improvement of the banana '*Musa acuminata*' reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genom*. 17:243. doi: 10.1186/s12864-016-2579-4

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res*. 17, 240–248. doi: 10.1101/gr.5681207

Minio, A., Lin, J., Gaut, B. S., and Cantu, D. (2017). How single molecule realtime sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci*. 8:826. doi: 10.3389/fpls.2017.00826

Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A., and Cantu, D. (2019). Diploid genome assembly of the wine grape Carménère. *G3* 9, 1331–1337. doi: 10.1534/g3.119.400030

Roach, M. J., Johnson, D. L., Bohlmann, J., van Vuuren, H. J. J., Jones, S. J. M., Pretorius, I. S., et al. (2018). Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet*. 14:e1007807. doi: 10.1371/journal.pgen.1007807

Song, P., Yang, J., Wang, C., Lu, Q., Shi, L., Tayler, S., et al. (2021). *Arabidopsis* N^6-methyladenosine reader CPSF30-L recognizes FUE signals to control polyadenylation site choice in liquid-like nuclear bodies. *Mol. Plant* 14:571–587. doi: 10.1016/j.molp.2021.01.014

Tu, M., and Li, Y. (2020). Profiling alternative 3'untranslated regions in sorghum using RNA-seq data. *Front. Genet*. 11:556749. doi: 10.3389/fgene.2020.556749

Wang, Z., Miao, H., Liu, J., Xu, B., Yao, X., Xu, C., et al. (2019). *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nat. Plants* 5, 810–821. doi: 10.1038/s41477-019-0452-6

Yin, D., Ji, C., Ma, X., Li, H., Zhang, W., Li, S., et al. (2018). Genome of an allotetraploid wild peanut *Arachis monticola*: a *de novo* assembly. *GigaScience* 7:giy066. doi: 10.1093/gigascience/giy066

Yin, D., Ji, C., Song, Q., Zhang, W., Zhang, X., Zhao, K., et al. (2019). Comparison of *Arachis monticola* with diploid and cultivated tetraploid genomes reveals asymmetric subgenome evolution and improvement of peanut. *Adv. Sci.* 28:1901672. doi: 10.1002/advs.201901672

Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., et al. (2013). Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biol*. 13:141. doi: 10.1186/1471-2229-13-141

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet*. 51, 865–876. doi: 10.1038/s41588-019-0402-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 December 2019
doi: 10.3389/fgene.2019.01223

Towards Improved Molecular Identification Tools in Fine Fescue (*Festuca* L., Poaceae) Turfgrasses: Nuclear Genome Size, Ploidy, and Chloroplast Genome Sequencing

Yinjie Qiu[1]*, Cory D. Hirsch[2], Ya Yang[3] and Eric Watkins[1]

[1] Department of Horticultural Science, University of Minnesota, St. Paul, MN, United States, [2] Department of Plant Pathology, University of Minnesota, St. Paul, MN, United States, [3] Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN, United States

Fine fescues (*Festuca* L., Poaceae) are turfgrass species that perform well in low-input environments. Based on morphological characteristics, the most commonly-utilized fine fescues are divided into five taxa: three are subspecies within *F. rubra* L. and the remaining two are treated as species within the *F. ovina* L. complex. Morphologically, these five taxa are very similar; both identification and classification of fine fescues remain challenging. In an effort to develop identification methods for fescues, we used flow cytometry to estimate genome size and ploidy level and sequenced the chloroplast genome of all five taxa. Fine fescue chloroplast genome sizes ranged from 133,331 to 133,841 bp and contained 113–114 genes. Phylogenetic relationship reconstruction using whole chloroplast genome sequences agreed with previous work based on morphology. Comparative genomics suggested unique repeat signatures for each fine fescue taxon that could potentially be used for marker development for taxon identification.

Keywords: fine fescue, chloroplast genome, phylogeny, comparative genomics, low input turfgrass

OPEN ACCESS

Edited by:
Wenqin Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Yingjuan Su,
Sun Yat-sen University, China
David J. Winter,
Massey University, New Zealand

****Correspondence:***
Yinjie Qiu
qiuxx221@umn.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *15 August 2019*
Accepted: *05 November 2019*
Published: *06 December 2019*

Citation:
Qiu Y, Hirsch CD, Yang Y and Watkins E (2019) Towards Improved Molecular Identification Tools in Fine Fescue (Festuca L., Poaceae) Turfgrasses: Nuclear Genome Size, Ploidy, and Chloroplast Genome Sequencing. Front. Genet. 10:1223. doi: 10.3389/fgene.2019.01223

INTRODUCTION

With ca. 450 species, Fescues (*Festuca* L., Poaceae) is a large and diverse genus of perennial grasses (Clayton and Renvoize, 1986). Fescue species are distributed mostly in temperate zones of both the northern and southern hemispheres, but most commonly found in the northern hemisphere (Jenkin, 1959). Several of the fescue species have been commonly used as turfgrass. Based on both leaf morphology and nuclear ITS sequences, fescue species can be divided into two groups: broad-leaved fescues and fine-leaved fescues (Torrecilla and Catalán, 2002). Broad-leaved fescues commonly used as turfgrass include tall fescue (*Festuca arundinacea* Schreb.) and meadow fescue (*Festuca pratensis* Huds.). Fine-leaved fescues are a group of cool-season grasses that include five commonly used taxa called fine fescues. Fine fescues include hard fescue (*Festuca brevipila* Tracey, 2n = 6x = 42), sheep fescue (*Festuca ovina* L., 2n = 4x = 28), strong creeping red fescue (*Festuca rubra* ssp. *rubra* 2n = 8x = 56), slender creeping red fescue [*F. rubra* ssp. *litoralis* (G. Mey.) Auquier 2n = 6x = 42], and Chewings fescue [*F. rubra* ssp. *fallax* (Thuill.) Nyman 2n = 6x = 42] (Ruemmele et al., 1995). All five taxa share very fine and narrow leaves and have been used for forage, turf, and ornamental purposes. They are highly tolerant to shade and drought and prefer low pH (5.5-6.5) and low fertility soils (Beard, 1972). Additionally, fine fescues grow well in the shade or sun, have reduced mowing requirements, and do not need additional fertilizer or supplemental irrigation (Ruemmele et al., 1995).

Based on morphological and cytological features, fine fescues are currently divided into two groups referred to as the *F. rubra* complex (includes *F. rubra* ssp. *litoralis*, *F. rubra* ssp. *rubra*, and *F. rubra* ssp. *fallax*) and the non-rhizomatous *F. ovina* complex (includes *F. brevipila* and *F. ovina*) (Ruemmele et al., 1995). While it is relatively easy to separate fine fescue taxa into their proper complex based on the presence and absence of rhizome, it is challenging to identify taxon within the same complex. In the *F. rubra* complex, both ssp. *litoralis* and ssp. *rubra* are rhizomatous, while ssp. *fallax* is non-rhizomatous. However, the separation of ssp. *litoralis* from ssp. *rubra* using rhizome length is challenging. Taxon identification within the *F. ovina* complex relies heavily on leaf characters; however, abundant morphological and ecotype diversity within *F. ovina* makes taxa identification difficult (Piper, 1906). This is further complicated by inconsistent identification methods between different continents. For example, in the United States, sheep fescue is described as having a bluish gray leaf color and hard fescue leaf blade color is considered green (Beard, 1972), while in Europe, it is the opposite (Hubbard, 1968). Because the ploidy level of the five taxa varies from tetraploid to octoploid, beyond morphological classifications, laser flow cytometry has been used to determine ploidy level of fine fescues and some other fescue species (Huff and Palazzo, 1998). A wide range of DNA contents within each complex suggests that the evolutionary history of each named species is complicated, and interspecific hybridization might interfere with species determination using this approach. Plant breeders have been working to improve fine fescues for turf use for several decades, with germplasm improvement efforts focused on disease resistance, traffic tolerance, and ability to perform well under heat stress (Casler, 2003). Turfgrass breeders have utilized germplasm collections from old turf areas as a source of germplasm (Bonos and Huff, 2013); however, confirming the taxon identity in these collections has been challenging. A combination of molecular markers and flow cytometry could be a valuable tool for breeders to identify fine fescue germplasm (Hebert et al., 2003).

Due to the complex polyploidy history of fine fescues, sequencing plastid genomes provides a more cost-effective tool for taxon identification than the nuclear genome because it is often maternally inherited, lacks heterozygosity, is present in high copies, and is usable even in partially degraded material (Bryan et al., 1999; Provan et al., 2001). Previous studies have developed universal polymerase chain reaction (PCR) primers to amplify non-coding polymorphic regions for DNA barcoding in plants for species identification (Baldwin et al., 1995; Demesure et al., 1995). However, the polymorphisms discovered from these regions are often single nucleotide polymorphisms that are difficult to apply using PCR screening methods. For these reasons, it would be helpful to assemble chloroplast genomes and identify simple sequence repeat (SSR) and tandem repeats polymorphisms. Chloroplast genome sequencing has been simplified due to improved sequencing technology. In turfgrass species, high throughput sequencing has been used to assemble the chloroplast genomes of perennial ryegrass (*Lolium perenne* cv. Cashel) (Diekmann et al., 2009), tall fescue (*Lolium arundinacea* cv. Schreb) (Cahoon et al., 2010), diploid *F. ovina*, *F. pratensis*, *Festuca altissima* (Hand et al., 2013), and bermudagrass (*Cynodon dactylon*) (Huang et al., 2017). To date, there is limited molecular biology information on fine fescue taxon identification and their phylogenetic position among other turfgrass species (Hand et al., 2013; Cheng et al., 2016). In this study, we used flow cytometry to confirm the ploidy level of five fine fescue cultivars, each representing one of the five commonly utilized fine fescue taxa. We then reported the complete chloroplast genome sequences of these five taxa and carried out comparative genomics and phylogenetic inference. Based on the genome sequence we identified unique genome features among fine fescue taxa and predicted taxon specific SSR and tandem repeat loci for molecular marker development.

RESULTS

Species Ploidy Level Confirmation

We used flow cytometry to estimate the ploidy levels of five fine fescue taxa by measuring the DNA content in each nucleus. DNA content was reflected by the flow cytometry mean PI-A value. Overall, fine fescue taxa had mean PI-A values roughly from 110 to 180 (**Figure 1** and **Figure S1**). *F. rubra* ssp. *rubra* cv. Navigator II (2n = 8x = 56) had the highest mean PI-A value (181.434, %rCV 4.4). *F. rubra* ssp. *litoralis* cv. Shoreline (2n = 6x = 42) and *F. rubra* ssp. *fallax* cv. Treazure II (2n = 6x = 42) had similar mean PI-A values of 137.852, %rCV 3.7 and 145.864, %rCV 3.5, respectively. *F. brevipila* cv. Beacon (2n = 6x = 42) had a mean PI-A of 165.25, %rCV 1.9, while *F. ovina* cv. Quatro (2n = 4x = 28) had a mean PI-A of 108.43, %rCV 2.9. Standard reference *L. perenne* cv. Artic Green (2n = 2x = 14) had a G1 phase mean PI-A of 63.91, %rCV 3.0. USDA *F. ovina* PI 230246 (2n = 2x = 14) had a G1 mean PI-A of 52.73 (histogram not shown). The estimated genome size of USDA PI 230246 was 4.67 pg/2C. Estimated ploidy level of *F. brevipila* cv. Beacon was 6.3, *F. ovina* cv. Quatro was 4.11, *F. rubra* ssp. *rubra* cv. Navigator II was 6.9, *F. rubra* ssp. *litoralis* cv. Shoreline was 5.2, and *F. rubra* ssp. *fallax* cv. Treazure II was 5.5 (**Table 1**). All newly estimated ploidy levels roughly correspond to previously reported ploidy levels based on chromosome counts.

Plastid Genome Assembly and Annotation of Five Fescue Taxa

A total of 47,843,878 reads were produced from the five fine fescue taxa. After Illumina adaptor removal, we obtained 47,837,438 reads. The assembled chloroplast genomes ranged from 133,331 to 133,841 bp. The large single copy (LSC) and small single copy (SSC) regions were similar in size between the sequenced fine fescue accessions (78 kb and 12 kb, respectively). *F. ovina* and *F. brevipila* in the *F. ovina* complex had exactly the same size inverted repeat (IR) region (42,476 bp). In the *F. rubra* complex, *F. rubra* ssp. *rubra* and *F. rubra* ssp. *litoralis* had the same IR size (21,235 bp). Species in the *F. rubra* complex had a larger chloroplast genome size compared to species in the *F. ovina*

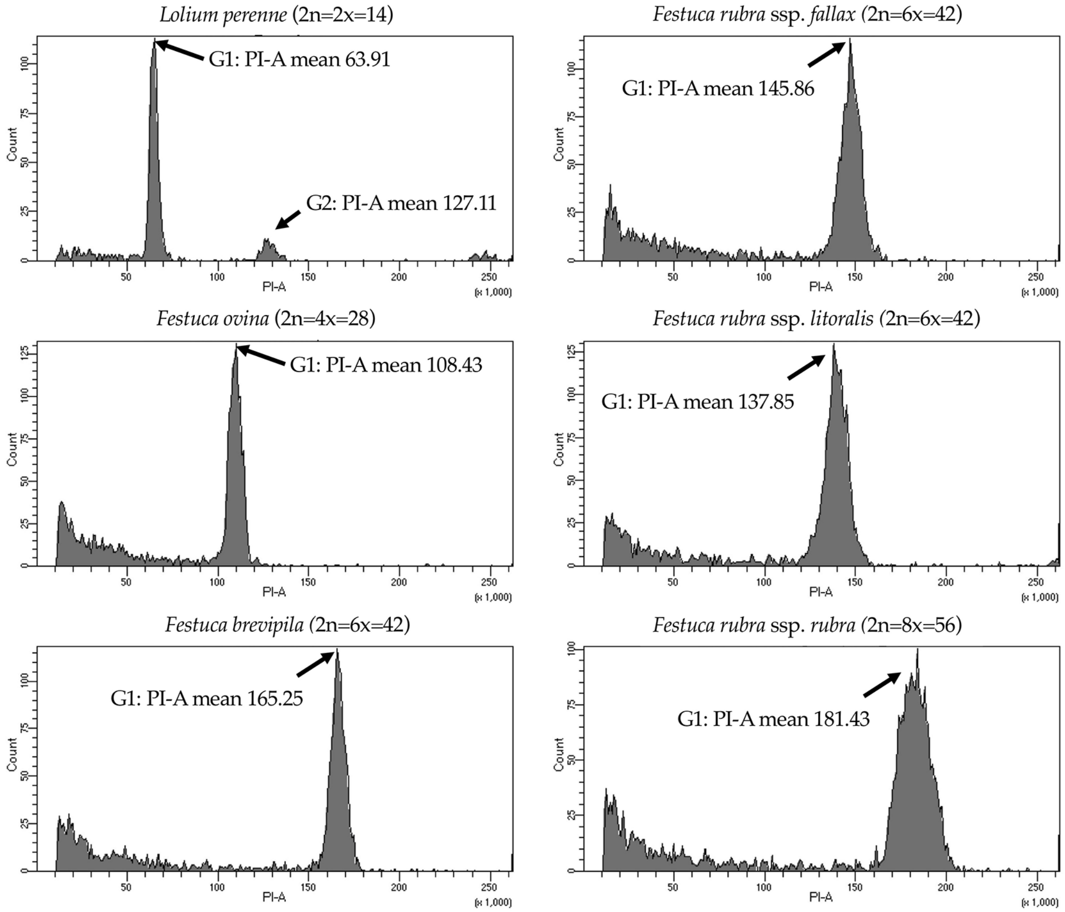


FIGURE 1 | Flow cytometry results for the five fine fescue taxa. *Lolium perenne* (2n = 2x = 14) was used as the reference. Flow cytometry was able to separate *F. rubra* ssp. *rubra* from the other two taxa in the *F. rubra* complex. The mean PI-A values of *F. rubra* ssp. *fallax* and *F. rubra* ssp. *litoralis* were similar (145.86 to 137.85).

TABLE 1 | Summary of flow cytometry statistics, genome size, and ploidy level estimation of fine fescue species. *Lolium perenne* 2C DNA content was used to calculate fine fescue and USDA *F. ovina* PI 230246 genome size, and calculated PI 230246 DNA content was used as reference to infer fine fescue ploidy level.

Species name	Chromosome count	Cultivar name	Mean PI-A	%rCV*	Estimated genome size (pg/nuclei)	Estimated ploidy level
F. brevipila	2n = 6x = 42	Beacon	165.3	1.9	14.6	6.3
F. ovina	2n = 4x = 28	Quatro	108.4	2.9	9.6	4.1
F. ovina PI 230246	2n = 2x = 14	NA	52.7	3.1	4.7	1.7
F. rubra ssp. *rubra*	2n = 8x = 56	Navigator II	181.4	4.4	16.1	6.9
F. rubra ssp. *litoralis*	2n = 6x = 42	Shoreline	137.9	3.7	12.2	5.2
F. rubra ssp. *fallax*	2n = 6x = 42	Treazure II	145.9	3.5	12.9	5.5
L. perenne	2n = 2x = 14	Artic Green	63.9	3.0	5.7	2.0

**%rCV: Quality of laser alignment. Low %rCV suggests high resolution sensitivity.*

complex. All chloroplast genomes shared similar GC content (38.4%) (**Figure 2**, **Table 2**). The fine fescue chloroplast genomes encoded for 113-114 genes, including 37 transfer RNAs (tRNA), 4 ribosomal RNAs (rRNA), and 72 protein-coding genes (**Table 2**). Genome structures were similar among all five fine fescue taxa sequenced, except that the pseudogene *accD* was annotated in all three subspecies of *F. rubra,* but not in the *F. ovina* complex (**Table S1**).

Chloroplast Genome IR Expansion and Contraction

Contraction and expansion of the IR regions resulted in the size variation of chloroplast genomes. We examined the four junctions in the chloroplast genomes, LSC/IRa, LSC/IRb, SSC/IRa, and SSC/IRb of the fine fescue and the model turfgrass species *L. perenne*. Although the chloroplast genome of fine fescue taxa was highly similar, some structural variations were

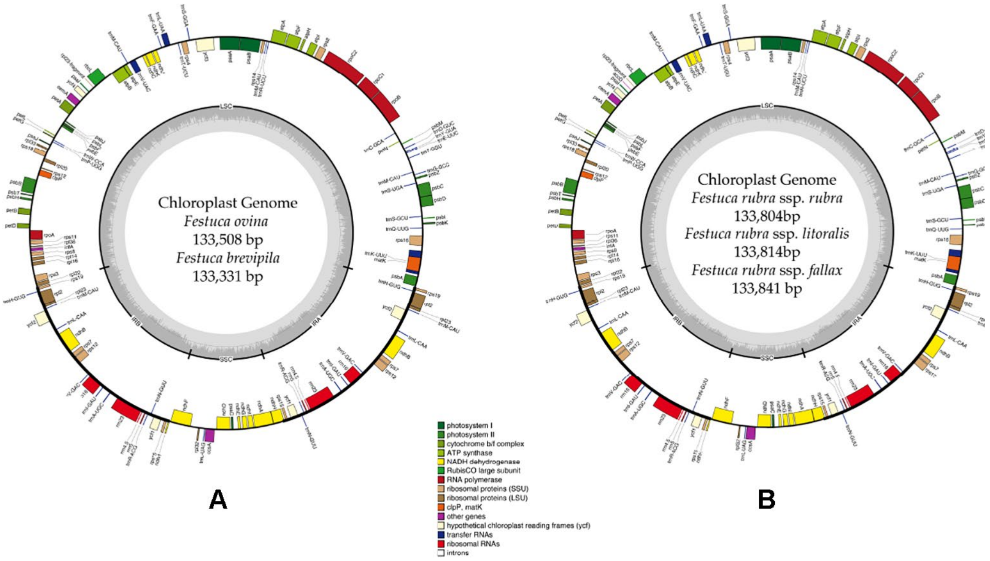


FIGURE 2 | Whole chloroplast genome structure of *F. ovina* complex **(A)** and *F. rubra* complex **(B)**. Genes inside the circle are transcribed clockwise, and genes outside are transcribed counter-clockwise. Genes belong to different functional groups are color coded. GC content is represented by the dark gray inner circle, and the light gray corresponded to the AT content. IRA(B), inverted repeat A(B); LSC, large single copy region; SSC, small single copy region.

TABLE 2 | Characteristics of fine fescue chloroplast genomes.

	***F. brevipila* cv. Beacon**	***F. ovina* cv. Quatro**	***F. rubra* ssp. *rubra* cv. Navigator II**	***F. rubra* ssp. *litoralis* cv. Shoreline**	***F. rubra* ssp. *fallax* cv. Treazure II**
NCBI GenBank ID	MN309822	MN309824	MN309825	MN309823	MN309826
Total Genome Size (bp)	133,331	133,508	133,804	133,814	133,841
Large Single Copy (bp)	78,462	78,632	78,888	78,909	78,882
Small Single Copy (bp)	12,393	12,400	12,446	12,435	12,451
Inverted Repeat (bp)	42,476	42,476	42,470	42,470	42,508
Ratio of LSC (%)	58.85	58.9	58.96	58.97	58.94
Ratio of SSC (%)	9.29	9.29	9.3	9.29	9.3
Ratio of IRs (%)	31.86	31.82	31.74	31.74	31.76
GC content (%)	38.4	38.4	38.4	38.4	38.4

still found in the IR/LSC and IR/SSC boundary. Similar to *L. perenne*, fine fescue taxa chloroplast genes *rpl22-rps19, rps19-psbA* were located in the junction of IR and LSC; *rps15-ndhF* and *ndhH-rps15* were located in the junction of IR/SSC. In the *F. ovina* complex, the *rps19* gene was 37 bp into the LSC/IRb boundary while in the *F. rubra* complex and *L. perenne*, the *rps19* gene was 36 bp into the LSC/IRb boundary (**Figure 3**). The *rsp15* gene was 308 bp from the IRb/SSC boundary in *F. ovina* complex, 307 bp in *F. rubra* complex, and 302 bp in *L. perenne*. Both the *ndhH* and the pseudogene fragment of the *ndhH* (*ψndhH)* genes spanned the junction of the IR/SSC. The *ψndhH* gene crossed the IRb/SSC boundary with 32 bp into SSC in *F. brevipila* and *F. ovina*, 9 bp in *F. rubra* ssp. *rubra* and *F. rubra* ssp. *litoralis*, 10 bp in *F. rubra* ssp. *fallax,* and 7 bp in *L. perenne*. The *ndhF* gene was 88 bp from the IRb/SSC boundary in *F. brevipila* and *F. ovina*, 91 bp in *F. rubra* ssp. *rubra*, 84 bp in *F. rubra* ssp. *litoralis*, 77 bp in *F. rubra* ssp. *fallax*, and 102 bp in *L. perenne*. Finally, the *psbA* gene was 87 bp apart from the IRa/LSC boundary into the LSC in *L. perenne* and *F. ovina* complex taxa but 83 bp in the *F. rubra* complex taxa.

Whole Chloroplast Genome Comparison and Repetitive Element Identification

Genome-wide comparison among five fine fescue taxa showed high sequence similarity with most variations located in intergenic regions (**Figure 4**). To develop markers for species screening, we predicted a total of 217 SSR markers for fine fescue taxa sequenced (*F. brevipila* 39; *F. ovina* 45; *F. rubra* ssp. *rubra* 45; *F. rubra* ssp. *litoralis* 46; *F. rubra* ssp. *fallax* 42) that included 17 different repeat types for the fine fescue species (**Figure 5**, **Table S2**). The most frequent repeat type was A/T repeats, followed by AT/AT. The pentamer AAATT/AATTT repeat was only presented in the rhizomatous *F. rubra* ssp. *litoralis* and *F. rubra* ssp. *rubra*, while ACCAT/ATGGT was only found in *F. ovina* complex species *F. brevipila* and *F. ovina*. Similar to SSR

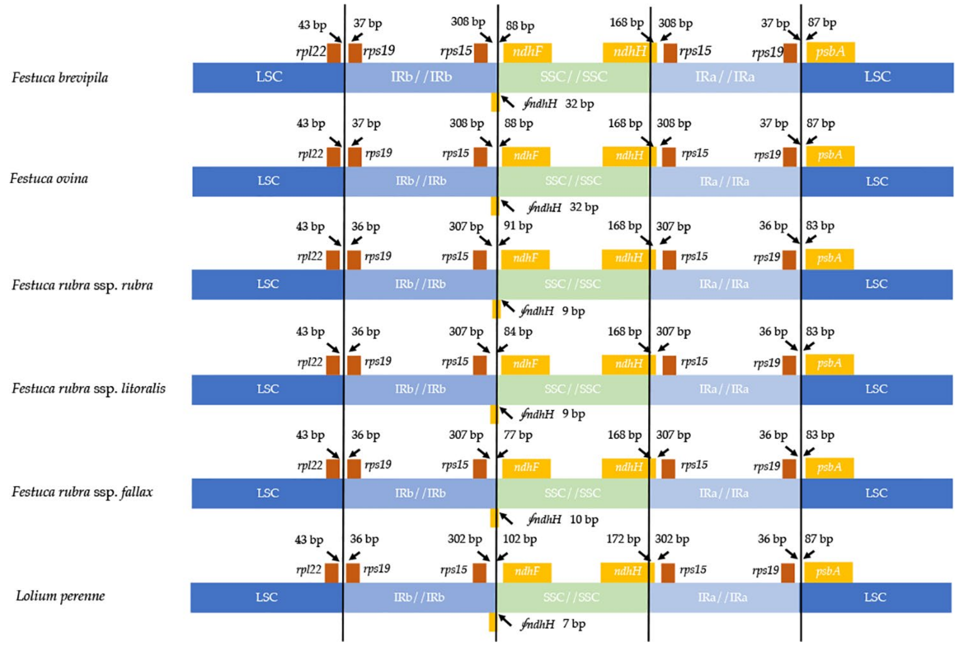


FIGURE 3 | Comparison for border positions of LSC, SSC, and IR regions among five fine fescues and *L. perenne*. Genes are denoted by boxes, and the gap between the genes and the boundaries is indicated by the number of bases unless the gene coincides with the boundary. Extensions of genes are also indicated above the boxes.

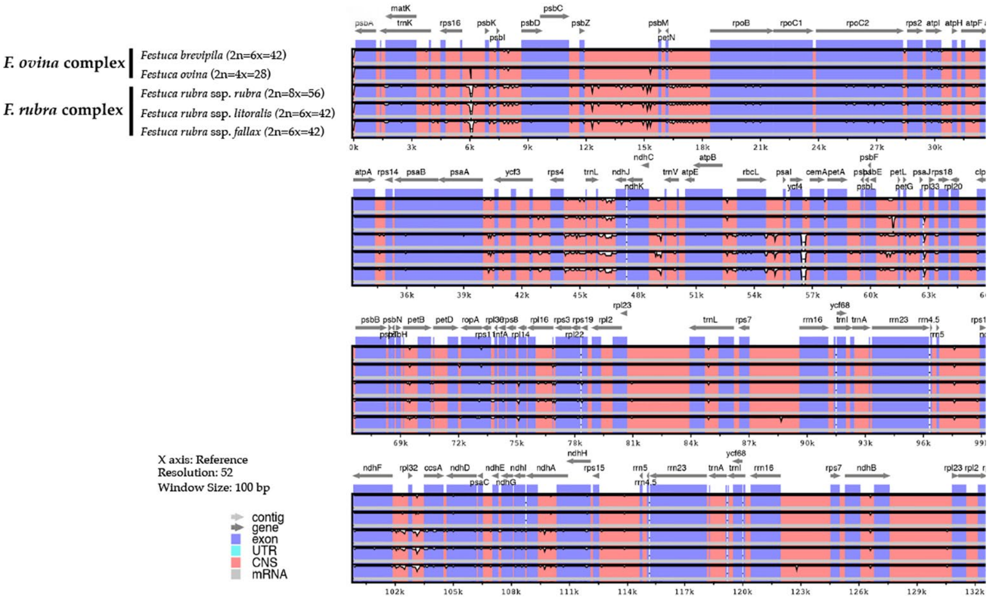


FIGURE 4 | Sequence identity plot of fine fescues chloroplast genome sequences with *F. ovina* (2x) as the reference using mVISTA. A cut-off of 70% identify was used for the plots, and the percent of identity varies from 50% to 100% as noted on the y-axis. Most of the sequence variations between fine fescues were in intergenic regions. Taxa in the *F. ovina* complex, *F. brevipila*, and *F. ovina* showed high sequence similarity. Similarly, subspecies within *F. rubra* complex also showed high sequence similarity.

loci prediction, we also predicted long repeats for the fine fescue species and identified a total of 171 repeated elements ranging in size from 20 to 51 bp (**Figure 5B**, **Table S3**). Complementary (C) matches were only identified in *F. brevipila* and *F. ovina*. *F. rubra* species had more palindromic (P) and reverse (R) matches. A number of forward (F) matches were similar between five taxa. Selected polymorphic regions were validated by PCR and gel electrophoresis assay (**Figure S2**).

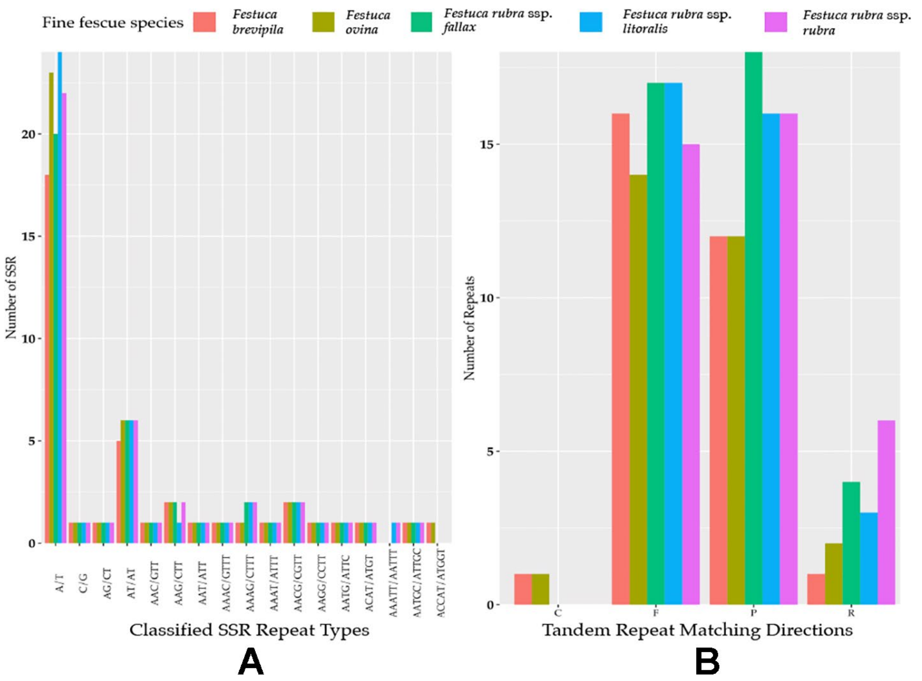


FIGURE 5 | (A) SSR repeat type and numbers in the five fine fescue taxa sequenced. Single nucleotide repeat type has the highest frequency. No hexanucleotide repeats were identified in the fine fescue chloroplast genomes sequenced. One penta-nucleotide repeat type (AAATT/AATTT) is unique to *F. rubra* ssp. *rubra* and *F. rubra* ssp. *litoralis*. One penta-nucleotide repeat type (ACCAT/ATGGT) is unique to *F. brevipila* and *F. ovina*. **(B)** Long repeats sequences in fine fescue chloroplast genomes. Complement (C) match was only identified in the *F. ovina* complex. Reverse (R) match has the most number variation in fine fescues.

SNP and InDel Distribution in the Coding Sequence of Five Fine Fescue Species

To identify single nucleotide polymorphisms (SNPs, non-reference allele in this content), we used the diploid *F. ovina* chloroplast genome (JX871940.1) as the reference for the mapping and used the genome annotation file to identify genic and non-genic SNPs. The total genic and non-genic sequence of (JX871940.1) were 60,582 and 72,583 bp, respectively. We found SNP polymorphisms were over-present within intergenic regions in the *F. ovina* complex (~0.3 SNP/Kbp more), while were under-present in the *F. rubra* complex (~0.5 SNP/Kbp less). Most InDels were located in intergenic regions of the fine fescue species (**Table 3**). Between *F. ovina* and the *F. rubra* complex, *the ropC2* gene had the most SNPs (4 vs 31). The *rbcL* gene also has a high level of variation (1 vs 14.3). In addition, *rpoB*, *ccsA*, NADH dehydrogenase subunit genes (*ndhH*, *ndhF*, *ndhA*), and ATPase subunit genes (*atpA*, *atpB*, *aptF*) also showed variation between *F. ovina* and *F. rubra* complexes. Less SNP and InDel variation were found within each complex (**Table 3**, **Tables S4 and S5**).

Nucleotide Diversity Calculation

A sliding window analysis successfully detected highly variable regions in the fine fescue chloroplast genomes (**Figure 6**, **Table S6**). The average nucleotide diversity (Pi) among fine fescue taxa was relatively low (0.00318). The IR region showed lower variability than the LSC and SSC region. There were several

TABLE 3 | Distribution of SNPs and InDels for the five fine fescue taxa sequenced in this study.

	F. brevipila	*F. ovina*	*F. rubra* ssp. *rubra*	*F. rubra* ssp. *litoralis*	*F. rubra* ssp. *fallax*
Total number of SNPs	98	134	638	615	624
SNPs in the coding region	35	52	306	301	300
SNPs in intergenic region	63	82	332	314	324
SNPs per Kbp in genic region	0.5777	0.8583	5.0510	4.9685	4.9520
SNPs per Kbp in non-genic region	0.8680	1.1297	4.5741	4.3261	4.4639
Total number of InDels	112	102	149	156	149
InDels in the coding region	22	17	27	26	27
InDels in intergenic region	90	85	122	130	122
Percentage of InDels in the intergenic region	80.36	83.33	81.88	83.33	81.88
Average sequencing depth	171.61	86.81	101.58	77.04	50.94

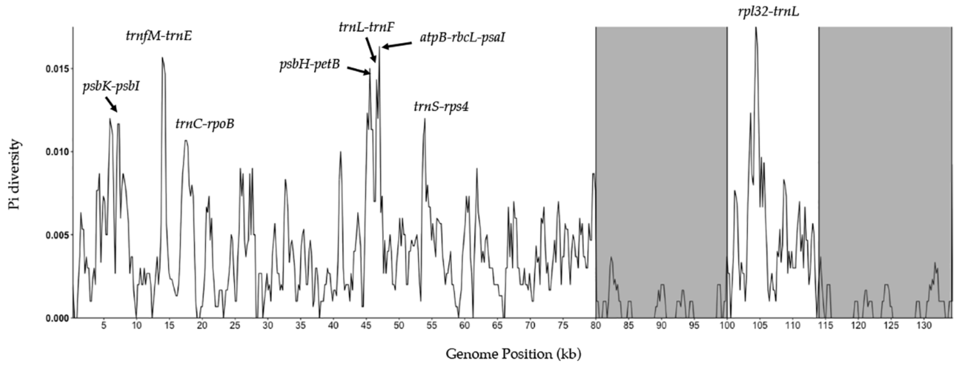


FIGURE 6 | Sliding window analysis of fine fescue whole chloroplast genomes. Window size: 600 bp, step size: 200 bp. X-axis: the position of the midpoint of a window (kb). Y-axis: nucleotide diversity of each window. Inverted repeat regions are highlighted in gray. *rpl32-trnL* region has the most nucleotide diversity followed by *psbH- petB-trnL-trnF-trnS-rps4* region.

divergent loci having a Pi value greater than 0.01 (*psbK-psbI*, *trnfM-trnE*, *trnC-rpoB*, *psbH-petB*, *trnL-trnF*, *trnS-rps4*, *aptB-rbcL-psaI*, and *rpl32-trnL*), but mostly within intergenic regions. The *rbcL-psaI* region contained a highly variable *accD-like* region in some fine fescue taxa, so we looked at the structural variation of 10 taxa in the *Festuca-Lolium* complex. We found taxa in broad-leaved fescue and *F. rubra* complex had similar structure, while *F. ovina* (2x, 4x) and *F. brevipila* had a 276 bp deletion in the *rbcL-psaI* intergenic region (**Figure 7**).

Phylogenetic Reconstruction of Fine Fescue Species

We reconstructed the phylogenetic relationships of taxa within the *Festuca-Lolium* complex using the chloroplast genomes sequenced in our study and eight publicly available complete chloroplast genomes including six taxa within the *Festuca-Lolium* complex (**Figure 8**). The dataset included 125,824 aligned characters, of which 3,923 were parsimony-informative and 91.11% characters are constant. The five fine fescue taxa were split into two clades ([ML]BS = 100). In the *F. ovina* complex, two *F. ovina* accessions included in the phylogenetic analysis, a diploid one from GenBank, and a tetraploid one newly sequenced in this study are paraphyletic to *F. brevipila* ([ML] BS = 100). All three subspecies of *F. rubra* formed a strongly supported clade ([ML]BS = 100). Together they are sisters to the *F. ovina* complex ([ML]BS = 100).

DISCUSSION

In this study, we used flow cytometry to determine the ploidy level of five fine fescue cultivars, assembled the chloroplast genomes for each, and identified structural variation and mutation hotspots. We also identified candidate loci for marker development to facilitate fine fescue species identification. Additionally, we reconstructed the phylogenetic relationships of the *Festuca-Lolium* complex using plastid genome information

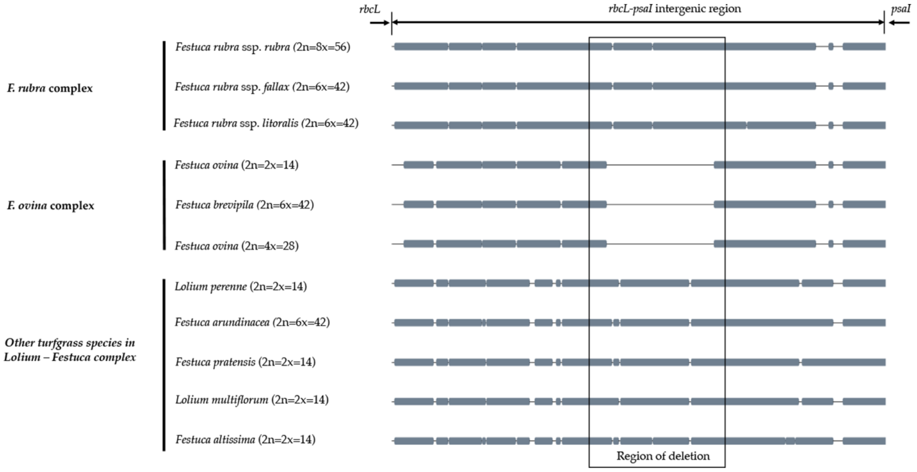

rbcL
rbcL-psaI intergenic region
psaI
F. rubra complex
Festuca rubra ssp. rubra (2n=8x=56)
Festuca rubra ssp. fallax (2n=6x=42)
Festuca rubra ssp. litoralis (2n=6x=42)
F. ovina complex
Festuca ovina (2n=2x=14)
Festuca brevipila (2n=6x=42)
Festuca ovina (2n=4x=28)
Other turfgrass species in Lolium – Festuca complex
Lolium perenne (2n=2x=14)
Festuca arundinacea (2n=6x=42)
Festuca pratensis (2n=2x=14)
Lolium multiflorum (2n=2x=14)
Festuca altissima (2n=2x=14)
Region of deletion

FIGURE 7 | The alignment of *rbcL-psaI* intergenic sequence shows that the pseudogene *accD* is missing in both *F. ovina* (2x, 4x) and *F. brevipila* but present in the *F. rubra* complex and other species examined in this study. Species were ordered by complexes.

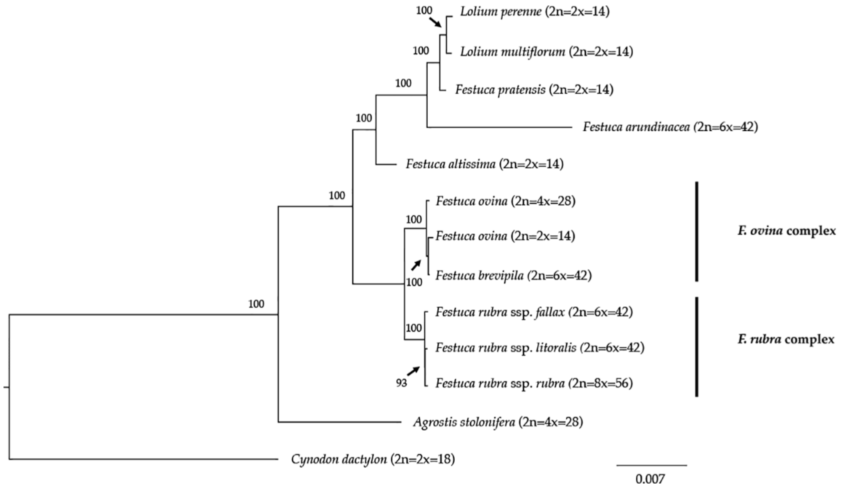

100
Lolium perenne (2n=2x=14)
100
Lolium multiflorum (2n=2x=14)
100
Festuca pratensis (2n=2x=14)
100
Festuca arundinacea (2n=6x=42)
Festuca altissima (2n=2x=14)
100
Festuca ovina (2n=4x=28)
100
Festuca ovina (2n=2x=14)
Festuca brevipila (2n=6x=42)
100
F. ovina complex
100
Festuca rubra ssp. fallax (2n=6x=42)
100
Festuca rubra ssp. litoralis (2n=6x=42)
93
Festuca rubra ssp. rubra (2n=8x=56)
F. rubra complex
Agrostis stolonifera (2n=4x=28)
Cynodon dactylon (2n=2x=18)
0.007

FIGURE 8 | Maximum likelihood (ML) phylogram of the *Festuca-Lolium* complex with 1,000 bootstrap replicates. Fine fescues were grouped into previous named complexes (*F. ovina* and *F. rubra*), sister to broad leaved fescues in the *Festuca-Lolium* complex.

generated in this study along with other publicly-available plastid genomes.

While most crop plants are highly distinctive from their close relatives, *Festuca* is a species-rich genus that contains species with highly similar morphology and different ploidy level. Consequently, it is difficult for researchers to interpret species identity. In our case, flow cytometry was able to successfully separate fine fescue taxa *F. brevipila* cv. Beacon, *F. ovina* cv. Quatro and *F. rubra* ssp. *rubra* cv. Navigator II based on the estimated ploidy levels. However, it is difficult to distinguish between *F. rubra* ssp. *litoralis* cv. Shoreline and *F. rubra* ssp. *fallax* cv. Treazure II as they had similar PI-A values based on flow cytometry.

We noticed that the average mean PI-A of the diploid *L. perenne* (63.91) was higher than the mean PI-A of diploid *F. ovina* (52.73), suggesting that *F. ovina* taxa have smaller genome size than *L. perenne*. The ploidy estimation in the *F. ovina* complex is fairly consistent, while the estimations of genome sizes in the *F. rubra* complex are smaller than we expected, even though these two complexes are closely related. Indeed, a similar finding was reported by Huff et al. (Huff and Palazzo, 1998) who reported that *F. brevipila* has a larger genome size than *F. rubra* ssp. *litoralis* and *F. rubra* ssp. *fallax*, both of which have the same ploidy level as *F. brevipila*. The range of variation in DNA content within each complex suggest a complicated evolutionary history in addition to polyploidization (Huff and Palazzo, 1998).

When we cannot identify taxon based on the ploidy level, we need different approaches to identify them. The presence and absence of rhizome formation could be taken into consideration; for example, *F. rubra* ssp. *fallax* cv. Treazure II is a bunch type turfgrass, while *F. rubra* ssp. *litoralis* cv. Shoreline forms short and slender rhizomes (Meyer and Funk, 1989). This method may not be effective because rhizome formation can be greatly affected by environmental conditions (Yang et al., 2015; Ma and Huang, 2016).

To further develop molecular tools to facilitate species identification, we carried out chloroplast genome sequencing. We assembled the complete chloroplast genomes of five low-input turfgrass fine fescues using Illumina sequencing. Overall, the chloroplast genomes had high sequence and structure similarity among all five fine fescue taxa sequenced, especially within each complex. All five chloroplast genomes share similar gene content except for the three species in the *F. rubra* complex that have a pseudogene Acetyl-coenzyme A carboxylase carboxyl transferase subunit (*accD*). The *accD* pseudogene is either partially or completely absent in some monocots. Instead, a nuclear-encoded ACC enzyme has been found to replace the plastic *accD* gene function in some angiosperm linage (Rousseau-Gueutin et al., 2013). Indeed, even though the *accD* pseudogene is missing in the *F. brevipila* chloroplast genome, the gene transcript was identified in a transcriptome sequencing dataset (unpublished data), suggesting that this gene has been translocated to nucleus genome. Previous studies have shown that broad-leaf fescues, *L. perenne*, *O. sativa*, and *H. vulgare* all had the pseudogene *accD* gene, while it was absent in diploid *F. ovina*, *Z. mays*, *S. bicolor*, *T. aestivum,* and *B. distachyon* (Hand et al., 2013). Broad-leaf and fine-leaf fescues diverged around 9 Mya ago (Fjellheim et al., 2006), which raises an interesting question about the mechanisms of the relocation of *accD* among closely related taxa in the *Festuca-Lolium* complex and even within fine fescue species.

In plants, chloroplast genomes are generally considered "single copy" and lack recombination due to maternal inheritance (Ebert and Peakall, 2009). For this reason, chloroplast genomes are convenient for developing genetic markers for distinguishing species and subspecies. We have identified a number of repeat signatures that are unique to a single species or species complex in fine fescue. For example, complement match is only identified in *F. ovina* complex, and *F. rubra* complex has more reversed matches. We also identified two SSR repeats unique to each of the two complexes. The first one consists of AAATT/AATTT repeat units is unique to *F. rubra* ssp. *litoralis* and *F. rubra* ssp. *rubra*, and the second one consists of ACCAT/ATGGT repeat units is unique to *F. brevipila* and *F. ovina*. In cases like the identification of hexaploids *F. brevipila*, *F. rubra* ssp. *fallax*, and *F. rubra* ssp. *litoralis*, it is critical to have these diagnostic repeats given all three taxa share similar PI-A values from flow cytometry. Taxon-specific tandem repeats could also aid the SSR repeats for species identification. We used chloroplast sequence developed candidate primer sets to solve the problem. Primer set (1) provided a clear separation of *F. rubra* ssp. *litoralis* cv. Shoreline and *F. rubra* ssp. *fallax* cv. Treazure II when flow cytometry was not able to separate them. Primer set (2) provided clear separation of *F. brevipila* cv. Beacon and *F. ovina* cv. Quatro, which provided an alternative method for *F. ovina* complex taxa identification. By combining both flow cytometry and candidate primer sets developed in this study, researchers will be able to identify fine fescue taxa within and between two complexes.

Nucleotide diversity analysis suggested that several variable genome regions exist among the five fine fescue taxa sequenced in this study. These variable regions included previously known highly variable chloroplast regions such as *trnL-trnF* and *rpl32-trnL* (Demesure et al., 1995; Dong et al., 2012). These regions have undergone rapid nucleotide substitution and are potentially informative molecular markers for characterization of fine fescue species.

Phylogeny inferred from DNA sequence is valuable for understanding the evolutionary context of a species. The phylogenetic relationship of fine fescue using whole plastid genome sequences agrees with previous classification based on genome size estimation and morphology (Huff and Palazzo, 1998; Cheng et al., 2016). The *F. ovina* complex includes *F. ovina* and *F. brevipila* and the *F. rubra* complex includes *F. rubra* ssp. *rubra*, *F. rubra* ssp. *litoralis* and *F. rubra* ssp. *fallax*, with the two rhizomatous subspecies (ssp. *rubra* and ssp. *literalis*) being sister to each other. Within the *Festuca-Lolium* complex, fine fescues are monophyletic and together sister to a clade consists of broad-leaved fescues and *Lolium*. In our analysis, *F. brevipila* (6x) is nested within *F. ovina* and sister to the diploid *F. ovina*. It is likely that *F. brevipila* arose from the hybridization between *F. ovina* (2x) and *F. ovina* (4x). Considering the complex evolutionary history of this genus, further research using nuclear loci sequences are needed to provide a more accurate phylogeny tree and validate this hypothesis.

The diversity of fine fescue provides valuable genetic diversity for breeding and cultivar development. Breeding fine fescue cultivars for better disease resistance, heat tolerance, and traffic tolerance could be achieved through screening wild accessions and by introgressing desired alleles into elite cultivars. Some work has been done using *Festuca* accessions in the USDA Germplasm Resources Information Network (GRIN) (https://www.ars-grin.gov) to breed for improved forage production in fescue species (Robbins et al., 2016). To date, there are 229 F. *ovina* and 486 F. *rubra* accessions in the USDA GRIN. To integrate this germplasm into breeding programs, plant breeders and other researchers need to confirm the ploidy level using flow cytometry and further identify them using molecular markers. Resources developed in this study could provide the tools to screen the germplasm accessions and refine the species identification so breeders can efficiently use these materials for breeding and genetics improvement of fine fescue species.

MATERIALS AND METHODS

Plant Material

Seeds from the fine fescue cultivars were obtained from the 2014 National Turfgrass Evaluation Program (www.ntep.org, USA) and planted in the Plant Growth Facility at the University of Minnesota, St. Paul campus under 16 h daylight (25°C) and 8 h dark (16°C) with weekly fertilization. Single genotypes of *F. brevipila* cv. Beacon, *F. rubra* ssp. *litoralis* cv. Shoreline, *F. rubra* ssp. *rubra* cv. Navigator II, *F. rubra* ssp. *fallax* cv. Treazure II, and *F. ovina* cv. Quatro were selected and used for chloroplast genome sequencing.

Flow Cytometry

To determine the ploidy level of the cultivars used for sequencing and compare them to previous work (2n = 4x = 28: *F. ovina*; 2n = 6x = 42: *F. rubra* ssp. *litoralis*, *F. rubra* ssp. *fallax*, and *F. brevipila*; 2n = 8x = 56: *F. rubra* ssp. *rubra*), flow cytometry was carried out using *Lolium perenne* cv. Artic Green (2n = 2x = 14) as the reference. Samples were prepared using CyStain PI Absolute P (Sysmex, product number 05-5022). Briefly, to prepare the staining solution for each sample, 12 µl propidium iodide (PI) was added to 12 ml of Cystain UV Precise P staining buffer with 6 µl RNase A. To prepare plant tissue, a total size of 0.5 cm x 0.5 cm leaf sample of the selected fine fescue was excised into small pieces using a razor blade in 500 µl CyStain UV Precise P extraction buffer and passed through a 50 µm size filter (Sysmex, product number 04-004-2327). The staining solution was added to the flow-through to stain nuclei in each sample. Samples were stored on ice before loading the flow cytometer. Flow cytometry was carried out using the BD LSRII H4760 (LSRII) instrument with PI laser detector using 480V with 2,000 events at the University of Minnesota Flow Cytometry Resource (UCRF). Data was visualized and analyzed on BD FACSDiva 8.0.1 software. To estimate the genome size, *L. perenne* DNA (5.66 pg/2C) was used as standard (Arumuganathan et al., 1999), and USDA PI 230246 (2n = 2x = 14) was used as diploid fine fescue relative (unpublished data). To infer fine fescues ploidy, estimation was done using equations (1) and (2) (Doležel et al., 2007).

$$\begin{aligned}&\text{Sample 2C DNA Content} =\\&\text{Standard 2C DNA Content}\left(\text{pg DNA}\right)\\&\times\frac{\left(\text{Sample G1 Peak Mean}\right)}{\left(\text{Standard G1 Peak Mean}\right)}\end{aligned}\qquad(1)$$

$$\text{Sample Ploidy} = \frac{2n \times \text{Sample pg/Nucleus}}{\text{Diploid Relative pg/Nucleus}}\qquad(2)$$

DNA Extraction and Sequencing

To extract DNA for chloroplast genome sequencing, 1 g of fresh leaves was collected from each genotype and DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega, USA) following manufacturer instructions. DNA quality was examined on 0.8% agarose gel and quantified *via* PicoGreen (Thermo Fisher, Catalog number: P11496). Sequencing libraries were constructed by NovoGene, Inc. (Davis, CA) using Nextera XT DNA Library Preparation Kit (Illumina) and sequenced in 150 bp paired-end mode, using the HiSeq X Ten platform (Illumina Inc., San Diego, CA, USA) with an average of 10 million reads per sample. All reads used in this study were deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA512126.

Chloroplast Genome Assembly and Annotation

Raw reads were trimmed of Illumina adaptor sequences using Trimmomatic (v. 0.32) (Bolger et al., 2014). Chloroplast genomes were *de novo* assembled using NovoPlasty v. 2.0 (Dierckxsens et al., 2016). Briefly, *rbcL* gene sequence from diploid *F. ovina* (NCBI accession number: JX871940) was extracted and used as the seed to initiate the assembly. NovoPlasty assembler configuration was set as follows: *k-mer* size = 39; insert size = auto; insert range = 1.8; and insert range strict 1.3. Reads with quality score above 25 were used to complete the guided assembly using *F. ovina* (NCBI accession number: JX871940) as the reference. Assembled plastid genomes for each taxon were manually corrected by inspecting the alignments of reads used in the assembly. The assembled chloroplast genomes were deposited under BioProject PRJNA512126, GenBank accession numbers MN309822-MN309826.

The assembled chloroplast genomes were annotated using the GeSeq pipeline (Tillich et al., 2017) and corrected using DOGMA online interface (https://dogma.ccbb.utexas.edu) (Wyman et al., 2004). BLAT [a BLAST-like alignment tool (Kent, 2002)] protein, tRNA, rRNA, and DNA search identity threshold was set at 80% in the GeSeq pipeline using the default reference database with the generate codon-based alignments option turned on. tRNAs were also predicted *via* tRNAscan-SE v2.0 and ARAGORN v 1.2.38 using the bacterial/plant chloroplast genetic code (Lowe and Eddy, 1997; Laslett and Canback, 2004). The final annotation was manually inspected and corrected using results from both pipelines. The circular chloroplast map was drawn by the OrganellarGenomeDRAW tool (OGDRAW) (Lohse et al., 2007).

Nucleotide Polymorphism of Fine Fescue Species

To identify genes with the most single nucleotide polymorphism, quality trimmed sequencing reads of the five fine fescues were mapped to the diploid *F. ovina* chloroplast genome (NCBI accession number: JX871940) using BWA v.0.7.17 (Li and Durbin, 2009). SNPs and short indels were identified using bcftools v.1.9 with the setting "mpileup -Ou" and called *via* bcftools using the -mv function (Quinlan and Hall, 2010). Raw SNPs were filtered using bcftools filter -s option to filter out SNPs with low quality (Phred score cutoff 20, coverage cutoff 20). The subsequent number of SNPs per gene and InDel number per gene was calculated using a custom perl script SNP_vcf_from_gene_gff.pl (https://github.com/qiuxx221/fine-fescue-).

To identify SSR markers for plant identification, MIcroSAtellite identification tool (MISA v 1.0) was used with a threshold of 10, 5, 4, 3, 3, and 3 repeat units for mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSRs, respectively (Thiel et al., 2003). The identification of repetitive sequences and structure of whole chloroplast genome was done *via* REPuter program online server (https://bibiserv.cebitec.uni-bielefeld.de/reputer) (Kurtz et al., 2001). Program configuration was set with minimal repeat size set as 20 bp and with sequence identify above 90%. Data was visualized using ggplot2 in R (v 3.5.3). Finally, the sliding window analysis was performed using DnaSP (v 5) with a window size of 600 bp, step size 200 bp to detected highly variable regions in the fine fescue chloroplast genome (Librado and Rozas, 2009).

Comparative Chloroplast Genomics Analysis

To compare fine fescue species chloroplast genome sequence variations, the five complete chloroplast genomes were aligned and visualized using mVISTA, an online suite of computation tools with LAGAN mode (Brudno et al., 2003; Frazer et al., 2004). The diploid *F. ovina* (NCBI accession number: JX871940) chloroplast genome and annotation were used as the template for the alignment.

Phylogenetic Analysis of Fine Fescues and Related Festuca Species

To construct the phylogenetic tree of the fine fescues using the whole chloroplast genome sequence, chloroplast genomes of eight species were downloaded from GenBank. Of the eight downloaded genomes, perennial ryegrass (*Lolium perenne*, AM777385), Italian ryegrass (*Lolium multiflorum*, JX871942), diploid *F. ovina* (JX871940), tall fescue (*Festuca arundiancea*, FJ466687), meadow fescue (*F. pratensis*, JX871941), and wood fescue (*Festuca altissima*, JX871939) were within the *Festuca-Lolium* complex. Turfgrass species outside of *Festuca-Lolium* complex including creeping bentgrass (*Agrostis stolonifera* L., EF115543) and *Cynodon dactylon* (KY024482.1) were used as an outgroup. All chloroplast genomes were aligned using the MAFFT program (v 7) (Katoh and Standley, 2013); alignments were inspected and manually adjusted. Maximum likelihood (ML) analyses was performed using the RAxML program (v 8.2.12) under GTR+G model with 1,000 bootstrap (Stamatakis, 2006). The phylogenetic tree was visualized using FigTree (v 1.4.3) (https://github.com/rambaut/figtree) (Rambaut, 2012).

CONCLUSIONS

Five newly-sequenced complete chloroplast genomes of fine fescue taxa were reported in this study. Chloroplast genome structure and gene contents are both conserved, with the presence and absence of *accD* pseudogene being the biggest structural variation between the *F. ovina* and the *F. rubra* complexes. We identified SSR repeats and long sequence repeats of fine fescues and discovered several unique repeats for marker development. The phylogenetic constructions of fine fescue species in the *Festuca-Lolium* complex suggested a robust and consistent relationship compared to the previous identification using flow cytometry. This information provided a reference for future fine fescue taxa identification.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI Bioproject PRJNA512126.

AUTHOR CONTRIBUTIONS

YQ performed the experiments, analyzed the data, and wrote the manuscript. CH helped analyze data, wrote perl scripts. YY helped with phylogenetic analysis. EW secured funding for this project, supervised this research, provided suggestions, and comments. All authors contributed to the revision of the manuscript and approved the final version.

FUNDING

This research is funded by the National Institute of Food and Agriculture, U.S. Department of Agriculture, Specialty Crop Research Initiative under award number 2017-51181-27222.

ACKNOWLEDGMENTS

The authors would like to thank minnesota supercomputing institute for the high-performance computing clusters. We would also like to thank Jill Ekar, Jason Motl, and Therese Martin for providing instruction on operating the flow cytometer.

This manuscript has been released as a pre-print at bioRxiv, (Qiu et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01223/full#supplementary-material

REFERENCES

Arumuganathan, K., Tallury, S., Fraser, M., Bruneau, A., and Qu, R. (1999). Nuclear DNA content of thirteen turfgrass species by flow cytometry. *Crop Sci.* 39, 1518–1521. doi: 10.2135/cropsci1999.3951518x

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., and Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Ann. Mo. Bot. Garden*, 82 (2), 247–277. doi: 10.2307/2399880

Beard, J. B. (1972). Turfgrass: Science and culture.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinform.* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bonos, S. A., and Huff, D. R. (2013). Cool-season grasses: Biology and breeding. Turfgrass. *Biology, Use Manage* 7, 591–660. doi: 10.2134/agronmonogr56.c17

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., et al. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721–731. doi: 10.1101/gr.926603

Bryan, G., Mcnicoll, J., Ramsay, G., Meyer, R., and De Jong, W. (1999). Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. *Theor. Appl. Genet.* 99, 859–867. doi: 10.1007/s001220051306

Cahoon, A. B., Sharpe, R. M., Mysayphonh, C., Thompson, E. J., Ward, A. D., and Lin, A. (2010). The complete chloroplast genome of tall fescue (Lolium arundinaceum; Poaceae) and comparison of whole plastomes from the family Poaceae. *Am. J. Bot.* 97, 49–58. doi: 10.3732/ajb.0900008

Casler, M. D. (2003). *Turfgrass biology, genetics, and breeding*. John Wiley & Sons.

Cheng, T., Xu, C., Lei, L., Li, C., Zhang, Y., and Zhou, S. (2016). Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Mol. Ecol. Resour.* 16, 138–149. doi: 10.1111/1755-0998.12438

Clayton, W. D., and Renvoize, S. A. (1986). Genera graminum. Grasses of the world. Genera graminum. Grasses of the World. 13.

Demesure, B., Sodzi, N., and Petit, R. (1995). A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol. Ecol.* 4, 129–134. doi: 10.1111/j.1365-294X.1995.tb00201.x

Diekmann, K., Hodkinson, T. R., Wolfe, K. H., Van Den Bekerom, R., Dix, P. J., and Barth, S. (2009). Complete chloroplast genome sequence of a major allogamous forage species, perennial ryegrass (Lolium perenne L.). *DNA Res.* 16, 165–176. doi: 10.1093/dnares/dsp008

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18–e18. doi: 10.1093/nar/gkw955

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233. doi: 10.1038/nprot.2007.310

Dong, W., Liu, J., Yu, J., Wang, L., and Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PloS One* 7, e35071. doi: 10.1371/journal.pone.0035071

Ebert, D., and Peakall, R. (2009). Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Mol. Ecol. Resour.* 9, 673–690. doi: 10.1111/j.1755-0998.2008.02319.x

Fjellheim, S., Rognli, O. A., Fosnes, K., and Brochmann, C. (2006). Phylogeographical history of the widespread meadow fescue (Festuca pratensis Huds.) inferred from chloroplast DNA sequences. *J. Biogeography* 33, 1470–1478. doi: 10.1111/j.1365-2699.2006.01521.x

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Hand, M. L., Spangenberg, G. C., Forster, J. W., and Cogan, N. O. (2013). Plastome sequence determination and comparative analysis for members of the Lolium-Festuca grass species complex. G3. *Genes Genomes Genet.* 3 (4), 607–616. doi: 10.1534/g3.112.005264

Hebert, P. D., Cywinska, A., Ball, S. L., and Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. London Ser. B.: Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Huang, Y.-Y., Cho, S.-T., Haryono, M., and Kuo, C.-H. (2017). Complete chloroplast genome sequence of common bermudagrass (Cynodon dactylon (L.) Pers.) and comparative analysis within the family Poaceae. *PloS One* 12, e0179055. doi: 10.1371/journal.pone.0179055

Hubbard, C. E. (1968). Grasses. A guide to their structure, identification, uses, and distribution in the British Isles. Grasses. A guide to their structure, identification, uses, and distribution in the British Isles.

Huff, D. R., and Palazzo, A. J. (1998). Fine fescue species determination by laser flow cytometry. *Crop Sci.* 38, 445–450. doi: 10.2135/cropsci1998.0011183X003800020029x

Jenkin, T. J. (1959). Fescue Species (Festuca L.). In: Roemer, T. & W. Rudorf. Handbuch der Pflanzenzüchtung, 418-434.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kent, W. J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi: 10.1093/nar/gkh152

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinform.* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinform.* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. doi: 10.1007/s00294-007-0161-y

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Ma, X., and Huang, B. (2016). Gibberellin-stimulation of rhizome elongation and differential GA-responsive proteomic changes in two grass species. *Front. In Plant Sci.* 7, 905. doi: 10.3389/fpls.2016.00905

Meyer, W. A., and Funk, C. R. (1989). Progress and Benefits to Humanity from Breeding Cool-Season Grasses for Turf 1. Contributions from breeding forage and turf grasses, 31-48.

Piper, C. V. (1906). *North American species of Festuca*. Department of Botany, Smithsonian Institution: US Government Printing Office. doi: 10.5962/bhl.title.53679

Provan, J., Powell, W., and Hollingsworth, P. M. (2001). Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends In Ecol. Evol.* 16, 142–147. doi: 10.1016/S0169-5347(00)02097-8

Qiu, Y., Hirsch, C. D., Yang, Y., and Watkins, E. (2019). Towards improved molecular identification tools in fine fescue (Festuca L., poaceae) turfgrasses: nuclear genome size, ploidy, and chloroplast genome sequencing. bioRxiv, 708149. doi: 10.1101/708149.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinform.* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rambaut, A. (2012). FigTree v1. 4.

Robbins, M. D., Staub, J. E., and Bushman, B. S. (2016). Development of fine-leaved Festuca grass populations identifies genetic resources having improved forage production with potential for wildfire control in the western United States. *Euphytica* 209, 377–393. doi: 10.1007/s10681-016-1644-z

Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., and Timmis, J. N. (2013). Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (accD) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiol.* 161, 1918–1929. doi: 10.1104/pp.113.214528

Ruemmele, B., Brilman, L., and Huff, D. (1995). Fine fescue germplasm diversity and vulnerability. *Crop Sci.* 35, 313–316. doi: 10.2135/cropsci1995.0011183X0 03500020003x

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinform.* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Thiel, T., Michalek, W., Varshney, R., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Torrecilla, P., and Catalán, P. (2002). Phylogeny of broad-leaved and fine-leaved Festuca lineages (Poaceae) based on nuclear ITS sequences. *Syst. Bot.* 27, 241–252. doi: 10.1043/0363-6445-27.2.241

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115. doi: 10.1093/nar/gks596

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinform.* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Yang, M., Zhu, L., Pan, C., Xu, L., Liu, Y., Ke, W., et al. (2015). Transcriptomic analysis of the regulation of rhizome formation in temperate and tropical lotus (Nelumbo nucifera). *Sci. Rep.* 5, 13059. doi: 10.1038/srep13059

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fgene.2020.00030

Varying Architecture of Heat Shock Elements Contributes to Distinct Magnitudes of Target Gene Expression and Diverged Biological Pathways in Heat Stress Response of Bread Wheat

*Peng Zhao, Sidra Javed, Xue Shi, Bingjin Wu, Dongzhi Zhang, Shengbao Xu and Xiaoming Wang**

State Key Laboratory of Crop Stress Biology for Arid Areas, College of Agronomy, Northwest A&F University, Yangling, China

OPEN ACCESS

Edited by:
Yin Li,
The State University of New Jersey,
United States

Reviewed by:
Ricardo Humberto Ramirez-Gonzalez,
John Innes Centre (JIC),
United Kingdom
Ting Zhao,
Zhejiang University, China

****Correspondence:***
Xiaoming Wang
wangxm@nwsuaf.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *18 September 2019*
Accepted: *09 January 2020*
Published: *04 February 2020*

Citation:
Zhao P, Javed S, Shi X, Wu B, Zhang D, Xu S and Wang X (2020) Varying Architecture of Heat Shock Elements Contributes to Distinct Magnitudes of Target Gene Expression and Diverged Biological Pathways in Heat Stress Response of Bread Wheat. Front. Genet. 11:30. doi: 10.3389/fgene.2020.00030

The heat shock transcription factor (HSF) binds to cis-regulatory motifs known as heat shock elements (HSEs) to mediate the transcriptional response of HSF target genes. However, the HSF–HSEs interaction is not clearly understood. Using the newly released genome reference sequence of bread wheat, we identified 39,478 HSEs (95.6% of which were non-canonical HSEs) and collapsed them into 30,604 wheat genes, accounting for 27.6% wheat genes. Using the intensively heat-responsive transcriptomes of wheat, we demonstrated that canonical HSEs have a higher propensity to induce a response in the closest downstream genes than non-canonical HSEs. However, the response magnitude induced by non-canonical HSEs was comparable to that induced by canonical HSEs. Significantly, some non-canonical HSEs that contain mismatched nucleotides at specific positions within HSEs had a larger response magnitude than that of canonical HSEs. Consistently, most of the HSEs identified in the promoter regions of heat shock proteins were non-canonical HSEs, suggesting an important role for these non-canonical HSEs. Lastly, distinct diverged biological processes were observed between genes containing different HSE types, suggesting that sequence variation in HSEs plays a key role in the evolution of heat responses and adaptation. Our results provide a new perspective to understand the regulatory network underlying heat responses.

Keywords: bread wheat, heat stress response, heat shock elements, heat shock transcription factor, heat-responsive transcriptomes

INTRODUCTION

Wheat (*Triticum aestivum* L.), a globally important crop, contributes about a fifth of the total calories consumed by humans (IWGSC, 2018). Increasing temperatures [heat stress (HS)], especially during the grain-filling stage, adversely affect the growth and development of wheat and causes a severe reduction in its yield and quality (Tack et al., 2015; Tao et al., 2015; Lesk et al., 2016; Zhao et al., 2017; Wang et al., 2018a). Thus, the identification of thermotolerant genes and the characterization of molecular mechanisms underlying HS responses and adaptations became urgent to improve wheat thermotolerance.

Heat shock proteins (HSPs), which were first identified based on their up-regulation during heat shock, play vital roles in HS responses by assisting in protein folding and preventing irreversible protein aggregation as chaperones (Waters, 2013). Under HS, heat shock factors (HSF), which converge the heat signaling transduced from several pathways and are regarded as the terminal link in heat signaling, bind to each other to form polymers and activate the expression of HSPs by recognizing and binding to conserved DNA sequences, known as heat shock elements (HSEs), in the promoter region of HSPs (von Koskull-Doring et al., 2007; Saidi et al., 2011; Bokszczanin et al., 2013; Vu et al., 2019). In many higher eukaryotes, HSFs are a diverged gene family, with family members varying in the stimuli needed for their activation, their affinity for HSEs, and the downstream targets (Takemori et al., 2009; Tian et al., 2010; Wang et al., 2018b). However, the factors that affect this stimulus and the corresponding activation efficiency are largely unknown.

The DNA binding domain of HSFs is highly conserved, implying that sequence variations in HSEs may be primarily responsible for the varying affinity in the HSF–HSEs interaction (Tian et al., 2010; Wang et al., 2018b). Canonical HSEs comprise at least three continuous inverted repeats of the pentanucleotide sequence, 5′-NGAAN-3′, alternating between 5′-NGAAN-3′ and 5′-NTTCN-3′, or vice versa, where N is any nucleotide. The “G” at the 2nd position of the 5′-NGAAN-3′ sequence and the “C” at the 4th position of the 5′-NTTCN-3′ sequence are key nucleotides and are the most critical for the binding of HSF. Each pentanucleotide sequence was defined as a subunit that was capable of binding one monomer of the HSF trimer. In *Drosophila*, the alternating subunits, the subunit number, and the position of HSEs within the promoter correlated with the HSF affinity and the magnitude of the HS response (Xiao et al., 1991; Fernandes et al., 1995; Tian et al., 2010).

Significantly, the 3rd and 4th positions of the 5′-NGAAN-3′ sequence and the 2nd and 3rd positions of the 5′-NTTCN-3′ sequence had lower effects on the HSF–HSEs interaction, allowing a mismatched nucleotide at these positions in the consensus sequence (Fernandes et al., 1994; Tian et al., 2010). The HSF affinity for HSEs also varies with HSE sequence variations (Santoro et al., 1998; Kremer and Gross, 2009; Jaeger et al., 2014). Besides, gapped HSEs, which contain an internal 5 bp block with little or no homology to the canonical motif flanked by canonical sequences in the proper orientation, also had a comparable binding affinity to the minimal functional HSE in *Drosophila*, as long as they exceed four subunits (Amin et al., 1988; Uffenbeck and Krebs, 2006; Tian et al., 2010). Generally, the architecture of HSEs is highly diverse in the genome of yeast, *Drosophila*, and humans, with deviations from the consensus sequence, orientation, and number of subunits that could influence the DNA binding affinity of HSF and the magnitude of target gene expression. In plants, a far larger and diverged HSF gene family was observed in our previous study (Wang et al., 2018b), implying a more complex HSF–HSEs interaction system. However, the distribution and architecture features of HSEs in the plant genome and their corresponding influence on the DNA binding affinity of HSF were not thoroughly investigated, hindering our understanding on the evolution of HSF–HSEs interactions that play key roles in the HS response.

Based on the newly released wheat genome reference sequence (IWGSC, 2018), we performed a genome-wide identification of HSEs that exhibited unexpectedly higher number of genes containing HSEs in promoter regions. Using our previous intensive time-course HS response transcriptomes of wheat flag leaves and filling grain (Wang et al., 2019), we elucidated the effects of position within the promoter and diverse architectures of HSEs on the magnitude of target gene expression under HS and showed that varying HSEs mediated different biological processes in the HS response. Our results provide a new perspective to understand the mechanisms and evolution of HS response and adaptation in plants.

MATERIALS AND METHODS

Identification and Extraction of Promoter Sequences of Wheat Genes

In order to obtain the promoter sequence of wheat high-confidence (HC) genes, we downloaded the reference genome sequence (IWGSC RefSeq v1.0, https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/iwgsc_refseqv1.0_all_chromosomes.zip), coding sequence of HC genes, IWGSC RefSeq v1.0 annotation, (https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_HighConf_CDS_2017Mar13.fa.zip) and genome annotation (IWGSC RefSeq v1.0 annotation, https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_HighConf_2017Mar13.gff3.zip) of the bread wheat cultivar Chinese Spring from IWGSC (IWGSC, 2018). Next, we screened the coding sequences (CDS) that started with “ATG” in genome annotation file, recorded their position on the genome, and extracted the upstream 2,000 bp sequence of CDS start sites as the promoter sequence (if the upstream sequences were less than 2,000 bp due to incomplete assembly, the longest upstream sequence was extracted) using bedtools (v2.27.1) (Quinlan and Hall, 2010) with the “getfasta” parameter. More details of each step and the relative scripts could been download on GitHub (https://github.com/biozhp/hse).

Genome-Wide Identification of Heat Shock Elements

At present, the definition of HSE structure in plants (Busch et al,., 2005; Guo et al,., 2008) and in *Drosophila* (Tian et al., 2010) are same. Furthermore, to investigate whether the HSE motif was conserved between *drosophila* and wheat, we downloaded the HSEs of *Drosophila* and Hymenoptera insects (ant, bee, aphid, etc.) from the published article (Tian et al., 2010; Nguyen et al., 2016) and retrieved the HSEs of *Arabidopsis thaliana* from

Abbreviations: HSEs, Heat shock elements; CDGs, Closest downstream genes; TTGs, Genes contained only one typical HSEs; TGGs, Genes contained only one gapped HSEs; TVGs, Genes contained only one varied HSEs; TSS, Transcription start site; MFH, Normalized maximal fold change.

JASPAR database (http://jaspar.genereg.net/). Then we performed the multiple motif alignment among HSE derived from above species and wheat with R package "MotifStack" (Ou et al., 2018) (**Supplemental Figures 1** and **2**). The results showed that the HSEs sequences of all species actually exhibited continuous inverted repeats of nGAAn as the HSE definition. More importantly, the HSEs in wheat and *A. thaliana* were located on different clades on the phylogenetic tree, and the HSEs of non-plant species did not clustered into single clade. These results demonstrated that the HSEs between plant and no-plant species were actually conserved, without obvious divergence. Therefore, we referred to the HSEs search procedure in *Drosophila* to identify HSE in wheat.

We developed a new search procedure referred to Tian et al. (2010), as following: (1) First, we identified typical HSEs comprising at least three continuous inverted repeats of the pentanucleotide sequence 5′-NGAAN-3′, alternating between 5′-NGAAN-3′ and 5′-NTTCN-3′ or vice versa, where N is any nucleotide; each pentanucleotide sequence was defined as a subunit. (2) For the sequence-varied HSEs, a total of one nucleotide was allowed to incur in a mismatch. (3) The nucleotides "G" and "C" in the subunit of 5′-NGAAN-3′ and 5′-NTTCN-3′, respectively, were key nucleotides. (4) The key nucleotides of the first and third subunits were not allowed to incur in a mismatch when the number of subunits was three, whereas the key bases were allowed to incur in a mismatch in every subunit when the number of subunits was more than three. Furthermore, we divided sequence-varied HSEs into gapped HSEs (contain a mismatched nucleotide at the "G" or "C" position of the middle subunit) and varied HSEs (contain a mismatched nucleotide that was not at the "G" or "C" position in the middle subunit). According to the criterion of HSEs, we designed a python program to identify HSEs in the promoters of wheat HC genes. The source scripts are available at GitHub (https://github.com/biozhp/hse).

Identification of Heat-Responsive Genes and Response Magnitude

In our previous study, wheat plants (*T. aestivum* cv. Chinese Spring) were first grown in a greenhouse under normal conditions and the plants at 15 days after anthesis were treated by heat stress (37°C) in growth chambers. The filling grain and flag leaves at 0, 5, 10, 30 min, 1, and 4 h under heat stress were harvested and subjected to 150 bp paired-end sequencing using the Illumina HiSeq X Ten platform. With the sample at 0 min time point which were not treated by HS as control, the differentially expressed genes (fold change ≥ 2.0 and false discovery rate-adjusted $p < 0.05$) at each heat stress treatment time point were identified (Wang et al., 2019). Genes that were differentially expressed at any HS treatment time point were defined as heat-responsive genes in each organism in the present analysis. Then, we investigated the maximal fold changes of heat-responsive genes among five time points (5, 10, 30 min, 1, and 4 h) under HS in each organism and normalized the maximal fold changes using the function of "scale (center = T, scale = T)" in R program. Finally, the normalized maximal fold change was used as an indicator to reflect the responsive magnitude of closest downstream genes (CDGs).

Gene Ontology Enrichment Analysis

The gene ontology (GO) annotation was obtained from our previous study (**Supplemental Table 1**) (https://zenodo.org/record/2541477/files/Genes_transcripts_FPKM.zip) (Wang et al., 2019). We used the R package "clusterProfiler" with the "enricher" function for enrichment analysis (Yu et al., 2012). The statistical significance of the GO enrichment was examined using the hypergeometric distribution test, followed by multiple-test correction using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). GO terms with $q < 0.01$ were retained for further analysis. The source codes and input files are available at GitHub (https://github.com/biozhp/hse/tree/master/example/enrichment).

Statistical Analysis

All statistical analyses were performed using the R-3.6.1. The function "chisq.test" with argument "correct = FALSE" was used for Pearson's Chi-squared test. One-way analysis of variance (one-way ANOVA) was performed with the function "aov" Multivariate analysis of variance (MANOVA) was performed with the function "manova" Multiple comparison analysis was performed with the R package "multcomp". The used source codes and input files are available at GitHub (https://github.com/biozhp/hse/blob/master/example/statistics.R).

Motif Enrichment Analysis

We performed motif enrichment analysis in the promoter regions of heat stress response genes and no heat stress response genes with the AME program (http://meme-suite.org/tools/ame) from MEME package (McLeay and Bailey, 2010), using the JASPAR CORE 2018 database as background. Then select motif with p-value < 0.01, q-value < 0.05 and e-value < 1e-5 for further analysis.

RESULTS

Identification of Heat Shock Elements in the Wheat Genome

Based on the fact that the HSE motif was conserved between *drosophila* and bread wheat (see "Methods"), we defined the wheat HSE identification criterion referring to the definition for HSEs in *Drosophila* (Tian et al., 2010). The genome of the bread wheat cultivar Chinese Spring exhibited 39,478 computationally identifiable HSEs in the promoter regions of all HC genes (**Supplemental Tables 2** and **5**), including 1,727 typical/canonical HSEs (three or more canonical 5 bp subunit sequences of 5′-NGAAN-3′ and 5′-NTTCN-3′ in alternation), 10,234 gapped HSEs (contain a mismatched nucleotide at the "G" or "C" position of the middle subunit), and 27,517 varied HSEs (contain a mismatched nucleotide that was not at the "G" or "C" position in the middle subunit) (**Supplemental Figure 3**). Unexpectedly, the number of varied and gapped HSEs was much larger than that of typical HSEs (**Figure 1A**), making it intriguing whether varied and gapped HSEs have the ability to interact with HSF to induce the expression of CDGs. Moreover, the subunit number of identified HSEs ranged from three to eight, and the number of all types of HSEs sharply decreased while the subunit number increased (**Figure 1A**). For the positions of HSEs in promoters, all HSE types were evenly distributed within the promoters (**Figure 1B**). Afterward, the distribution of HSEs among three wheat subgenomes was uneven, with the largest and smallest numbers of typical and varied HSEs on the A subgenome and the gapped HSEs on the D subgenome, respectively (**Figure 1C**). These results demonstrated that the wheat genome contains massive HSEs and exhibits evolutionary divergence for the subunit number, position, and sequence conformity.

The identified HSEs were collapsed into the promoter regions of 30,604 genes (27.62% of all HC genes annotated in IWGSC RefSeq 1.0). For these genes, 968, 5,816, and 16,906 genes contained only one typical HSE, gapped HSEs, or varied HSEs in their promoter regions and were designed as TTGs (Genes contained only one typical HSEs), TGGs (Genes contained only one gapped HSEs), and TVGs (Genes contained only one varied HSEs), respectively. The remaining 6,914 genes contained more than one HSEs in their promoter regions. The TTGs, TGGs, and TVGs were also unevenly distributed among three wheat subgenomes, with the largest number of TTGs, TGGs, and TVGs in the A subgenomes (**Figure 1D**). For clarity and accuracy, only TTGs, TGGs, and TVGs were used in the following analysis.

Varying Architecture of Heat Shock Elements Affect Heat Response of Closest Downstream Genes

To understand how the different types of HSEs affect the expression of CDGs, we comprehensively investigated the expression fold change of TTGs, TGGs, and TVGs under HS, which is a well-known inducing factor of HSF binding to HSEs; these, in turn, activate the expression of target genes, using the heat-responsive transcriptomes of wheat flag leaves and filling grain under HS at 0 min, 5 min, 10 min, 30 min, 1 h, and 4 h, as reported in our previous study (Wang et al., 2019). In filling grain, a total of 33 TTGs (3.4%), 75 TGGs (1.3%), and 339 TVGs (2.0%) responded to HS, compared to the untreated samples (fold change ≥ 2.0 and false discovery rate-adjusted $p < 0.05$), accounting for 1.6%, 3.5% and 15.9%, respectively, of the grain heat-responsive genes identified in our previously study. Similarly, in flag leaves, the relative numbers were 108 TTGs (11.2%), 301 TGGs (5.2%), and 1,212 TVGs (7.2%), accounting for 1.6%, 4.4% and 17.5%, respectively, of the heat-responsive genes in flag leaves. As expected, TTGs had a significantly higher

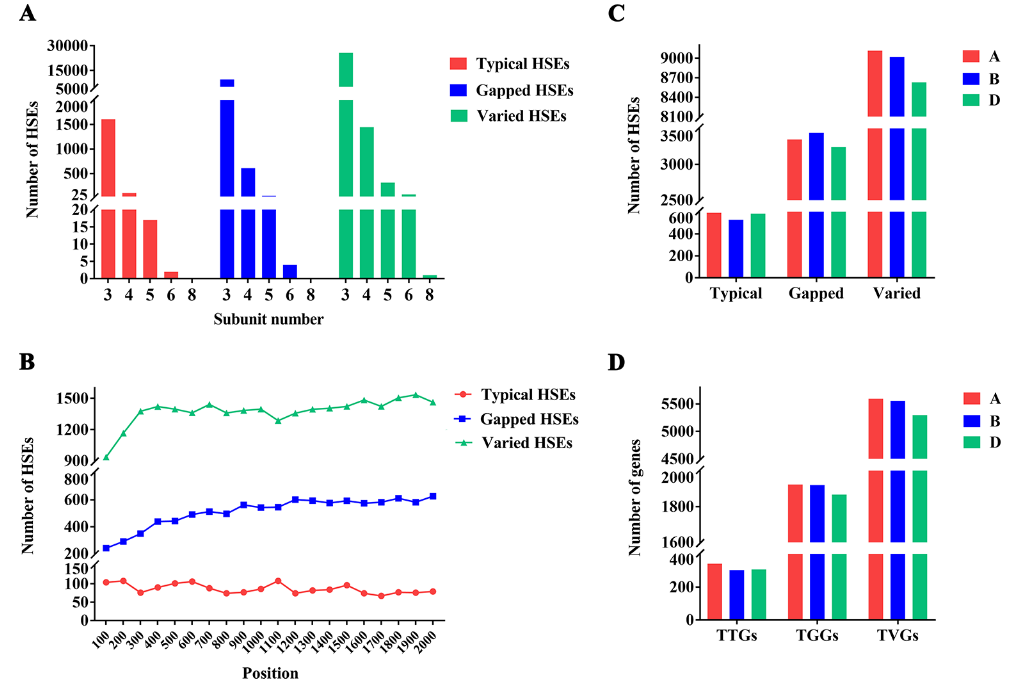


FIGURE 1 | Distribution of HSEs and its CDGs in wheat genome **(A)** Number of different HSE types. The x-axis represents the subunits and the y-axis represents the number of HSEs. Red, blue, and green represent the number of typical HSEs, gapped HSEs, and varied HSEs, respectively. **(B)** Positions of different types of HSEs within promoters. The x-axis represents the position. The "0" indicates the "TSS" and the "2000" indicates the upstream 2000th bp of TSS. The y-axis represents the number of HSEs. Red, blue, and green represent the number of typical HSEs, gapped HSEs, and varied HSEs, respectively. **(C)** Distribution of HSEs among three wheat subgenomes. The x-axis represents the different types of HSEs and the y-axis represents the number of HSEs. Red, blue, and green represent the **(A, B, D)** subgenome, respectively. **(D)** Distribution of genes containing different types of HSEs among three wheat subgenomes. The x-axis represents the genes containing different types of HSEs and the y-axis represents the number of genes. Red, blue, and green represent the A, B, and D subgenome, respectively.

proportion to response to HS than TGGs and TVGs in grain (Pearson's Chi-squared test, X^2 = 24.6, p = 4.5E-06) and in leaves (X^2 = 56.5, p = 5.5E-13), implying that the typical HSEs have higher abilities or affinities for binding to HSFs. These results demonstrated that the presence of HSEs was not equal to the HS response, and that HSEs sequence variations also affected this response, suggesting a more complex network and mechanism of HSF in mediating a HS response.

In detail, to further understand why some of the TTGs, TGGs, and TVGs do not respond to HS, we analyzed the effects of position within promoters and subunit number of HSEs on the response of CDGs to HS, using a MANOVA. The results showed that these two factors significantly affect the responses of TTGs in leaves and TVGs in grain and leaves (**Table 1**). Interestingly, the largest effects were observed on the response of TVGs in both grain and leaves, implying that position and subunit number of HSEs could compensate for the adverse effects of mismatched nucleotides in HSEs to some extent. Generally, a higher proximity of HSEs and TSS resulted in a higher ratio of CDGs that responded to HS, with the exception of TGGs (**Figure 2A**). For the subunit number, the HSE subunits four and five accounted for most of the HS responsiveness in HSEs (**Figure 2B**). The above results indicated that the architecture of HSEs affected whether CDGs respond to HS, suggesting that the HS response of a certain gene can be modulated by modifying the structure of HSEs in the promoter sequence.

Sequence-Varied Heat Shock Elements Had a Comparable Heat Stress Response Magnitude With Typical Heat Shock Elements

Given that the HSE architecture affects whether CDGs respond to HS, it is intriguing whether this architecture affects the HS responsive magnitude. To understand this relationship, we used the normalized maximal fold change (MFH) among five time points (5 min, 10 min, 30 min, 1 h, and 4 h) under HS in each organism as an indicator of the responsive magnitude of CDGs. The MFH of all HS-responsive TTGs, TGGs, and TVGs was first illustrated and no significant differences were observed among the three gene types in either grain (One-way ANOVA, p = 0.143) or leaves (p = 0.08) (**Supplemental Figure 4**), suggesting that the HS response magnitude of gapped and varied HSEs was comparable for CDGs with typical HSEs, although typical HSEs had a higher propensity to induce CDGs in response to HS. These results are consistent with the fact that almost all of the HSEs in the promoter regions of HSP genes, which are well-known target genes of HSF and marker genes in the HS response due to their sharply up-regulated expression (Waters, 2013), were varied HSEs (**Supplemental Figures 5–8**).

Furthermore, the relationship between HSE architecture (the position within promoter and the subunit number) and the responsive magnitude of each HSE type was investigated. Results showed that upon higher proximity of HSEs to TSS, a higher response magnitude of TVGs in grain, and TGGs and TVGs in leaves (**Tables 2** and **3**). The subunit number only significantly affects the responsive magnitude of TVGs in leaves, for which the four HSE subunits confer a higher response magnitude (**Supplemental Figure 9**). These results demonstrated that sequence variations in HSEs (HSEs type) did not affect the HS responsive magnitude of CDGs, whereas the HSEs architecture indeed contributes to this magnitude in each HSE type.

Mismatched Nucleotide At a Specific Position Within Heat Shock Elements Had a Larger Heat Stress Response Magnitude

Although no significantly different response magnitudes were observed among the three HSE types, whether mutations on a specific position or a specific mismatched nucleotide in HSE sequences correlated with the response magnitude is an open question. First, we excluded the effects of subunit number by only focusing on the three HSE subunits that had the largest numbers in our analysis. Then, using one-way analysis of covariance with the position of HSEs within the promoter as covariance, we found four positions within HSEs at which mismatched nucleotides significantly affected the response magnitude compared with that of typical HSEs (**Figure 3A**). For instance, the mismatch on the second or third positions in the second subunit of the sequence 5′-NTTCN-3′ in grain, and the 2nd position in the first subunit of the sequence 5′-NTTCN-3′ in leaves significantly affected the response magnitude. Unexpectedly, mismatches at the four positions conferred a significantly larger response magnitude than that of canonical HSEs (**Figure 3A**), suggesting evolutionary advantages of nucleotide mutations at these positions. To further support this conclusion, massive sequence varied HSEs at these four positions were observed in the promoter regions of HSP genes, especially the HSPs that had a larger response magnitude after HS

TABLE 1 | The effects of position within promoters and subunit number of HSEs in genes response to HS using multivariate analysis of variance.

	Grain			Leaves		
	TTGs	TGGs	TVGs	TTGs	TGGs	TVGs
F-value	7.697	3.705	25.589	25.315	0.857	72.407
P-value	4.83E-04**	2.47E-02**	8.01E-12**	1.93E-11**	4.25E-01	<2.2e-16**
Positions *P-value*	1.81E-04**	6.02E-01	7.07E-12**	3.19E-10**	8.96E-01	<2.2e-16**
Subunits *P-value*	2.74E-01	7.41E-03**	3.06E-02*	2.42E-03**	8.96E-01	4.00E-06**

** and ** represent the significant difference of* p-value *< 0.05 and* p-value *< 0.01, respectively.*

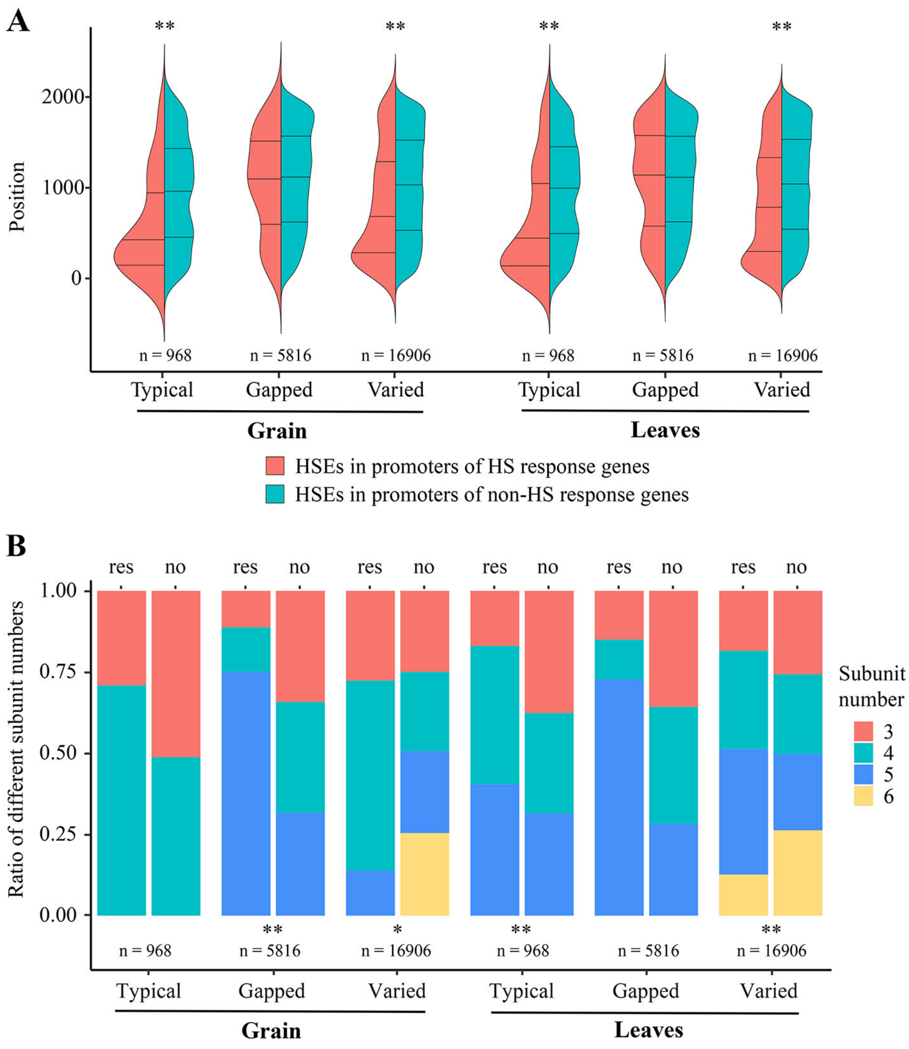


FIGURE 2 | Effects of varying architecture in different HSE types on the HS response of CDGs **(A)** HSE positions within promoters. The x-axis represents the different types of HSEs. The y-axis represents the positions of HSEs within promoters and the "0" indicates the "ATG". Red represents the HSEs in promoters of HS response genes, green represent the HSEs in promoters of non-HS response genes. **(B)** Ratio of different subunit numbers of HSEs. The x-axis represents the different types of HSEs and the y-axis represents the ratio of genes containing different subunit number of HSEs to all genes containing each type of HSEs. The "res" and "no" indicate the HS response gene set and the non-HS response gene set, respectively. Red, green, blue, and yellow represent the HSEs with 3, 4, 5, 6 subunits, respectively. * and ** represent the significant difference of p-value < 0.05 and p-value < 0.01, respectively.

treatment (**Supplemental Figures 5–8**). It is also noteworthy that these mismatched positions were different between grain and leaves, consistent with our previous results that these two organisms exploit different molecular mechanisms and networks underlying the HS response (Wang et al., 2019).

Furthermore, we analyzed the effects of different mismatched nucleotides on these four positions using multiple comparative analyses. Different magnitudes were observed among HSEs that had different mismatched nucleotides on the same position, although some difference was not statistically significant (**Figures 3B–E**). For example, the mismatched nucleotide "A" on the 2nd position of the second subunit in the sequence 5′-NTTCN-3′ in grain had a larger response magnitude than the mismatched nucleotides "G" and "C" at the same position. These results showed that mutations in HSE sequences have preferences for position and nucleotides in the evolution of the HS response and adaptation, providing clues for the response magnitude improvement of HSEs.

Distinctly Functional Divergence of Genes Containing Different Heat Shock Elements Types

Due to the contribution of HSE architecture to the HS response of CDGs, it is intriguing whether different HSE types were

TABLE 2 | The effects of position within promoters and subunit number of HSEs in genes response magnitude using analysis of covariance.

Tissues	Type	Variates	*F-value*	*P-value*
Grain	TTGs	Positions	0.478	4.95E-01
		Subunits	1.071	3.09E-01
	TGGs	Positions	0.155	6.95E-01
		Subunits	1.304	2.57E-01
	TVGs	Positions	5.744	1.71E-02*
		Subunits	1.910	1.68E-01
Leaves	TTGs	Positions	1.580	2.12E-01
		Subunits	0.018	8.94E-01
	TGGs	Positions	9.143	2.71E-03**
		Subunits	0.264	6.08E-01
	TVGs	Positions	26.850	2.58E-07**
		Subunits	12.040	5.38E-04**

** and ** represent the significant difference of* p-value *< 0.05 and* p-value *< 0.01, respectively.*

TABLE 3 | The effects of position within promoters in genes response magnitude.

Tissues	Type	*Coefficient of correlation*	*P-value*
Grain	TVGs	-0.138	1.12E-02*
Leaves	TGGs	-0.174	2.39E-03**
	TVGs	-0.158	3.07E-08**

** and ** represent the significant difference of* p-value *< 0.05 and* p-value *< 0.01, respectively.*

involved in different biological functions and pathways in the HS response. We performed a GO enrichment analysis for HS-response and non-HS-response of TTGs, TGGs, and TVGs, and for genes that do not contain HSEs in the promoter region but respond to HS. Interestingly, genes with different HSE types showed significantly distinct GO enrichment terms (**Figure 4**), suggesting a contribution of HSE sequence variations in the evolution of HS adaptation and response. The HS-responsive TTGs was mainly over-represented in the well-known HS response terms, such as "response to heat" and "chaperonemediated protein folding" whereas the HS-responsive TVGs showed significant over-representation in terms involved in extended HS response processes, such as "chaperone binding" "regulation of cell differentiation" and "positive regulation of response to oxidative stress". The HS-responsive TGGs was only enriched in terms of "response to heat" in grain, and had no significant enrichment terms in leaves. Combined with the previous result showing that TGGs had a lower propensity to induce an HS response of CDGs (when compared to TTGs and TVGs), we proposed that gapped HSEs tend to release from HSF regulation. It is also interesting that most of the over-represented terms of the HS-responsive genes, which contain no HSEs in their promoter regions but respond to HS, were not overlapped with terms that were enriched in genes containing HSEs; this suggests an important role for HSEs in the evolution of HS response and adaptation.

Unexpectedly, a number of GO terms were also significantly enriched in the non-HS-responsive TTGs, TGGs, and TVGs (**Figure 4**). For example, "oxidoreductase activity" and energy metabolism-related terms were enriched for non HS-responsive TVGs, and "mitochondrial ribosome," "mitochondrial small ribosomal subunit," and "organellar small ribosomal subunit" were enriched for non HS-responsive TTGs, posing an intriguing question as to what roles do these processes play in the evolution of HS response and adaptation. These results demonstrate that different HSE types derive distinctly diverged HS response processes and provides a new perspective for understanding the evolution of HS response and adaptation.

DISCUSSION

In this study, we comprehensively identified the distribution of HSEs and illustrated, for the first time, that varying HSE architecture affects the HSF DNA binding affinity and the corresponding response magnitude of CDGs in plants, thus mediating different HS response processes. Our results, including the large number of genes containing HSEs, the vast majority of varied HSEs, the comparable or higher HS response magnitude of genes containing varied HSEs, and the diverged HS response processes mediated by different HSEs types, suggest a complex interaction network in HS response and provide a new perspective to understand the HS response and adaptation.

HSEs that have mismatched nucleotides at specific positions have larger response magnitudes than that of typical HSEs. Furthermore, the diverged biological process that different HSE types are involved in, proposed an important role for HSE sequence variation in the evolution of HS response and adaptation in plants. It seems like differences in the HS response magnitude of specific genes and biological processes are the result of HSE architecture variation, instead of variations in DNA binding sequence of HSFs or the adjustment of the interacting proteins with HSFs. In our previous study, the oligomerization domain and the transcriptional activation domain of HSFs exhibited larger sequence divergences than that of DNA binding domains during plant evolution (Wang et al., 2018b), which may also contribute to the variations in HS response magnitude of target genes. Therefore, research aimed at the coevolution of HSEs and HSF will be vital for understanding the evolution of HS response and HS adaptation in plants.

It is an interesting question whether the varied HSEs were derived from mutations of typical HSE sequences or derived from evolution and natural selection of mutations from non-HSE sequences. Theoretically, the first hypothesis only needs one mutation, whereas the second hypothesis needs one or more mutations, making the first hypothesis more reasonable. However, in our results, the non-overlapped GO enrichment terms between genes containing typical HSEs and genes containing varied HSEs make the answer increasingly ambiguous. In the future, HS response analysis in more plant species, especially in ancient plants, will facilitate the answer to this question.

In our previous study, small HSPs, key genes in the HS response, maintained higher transcriptional levels in filling grains than in flag leaves (Wang et al., 2017). Furthermore, we also found that the number of HS-responsive genes, the response

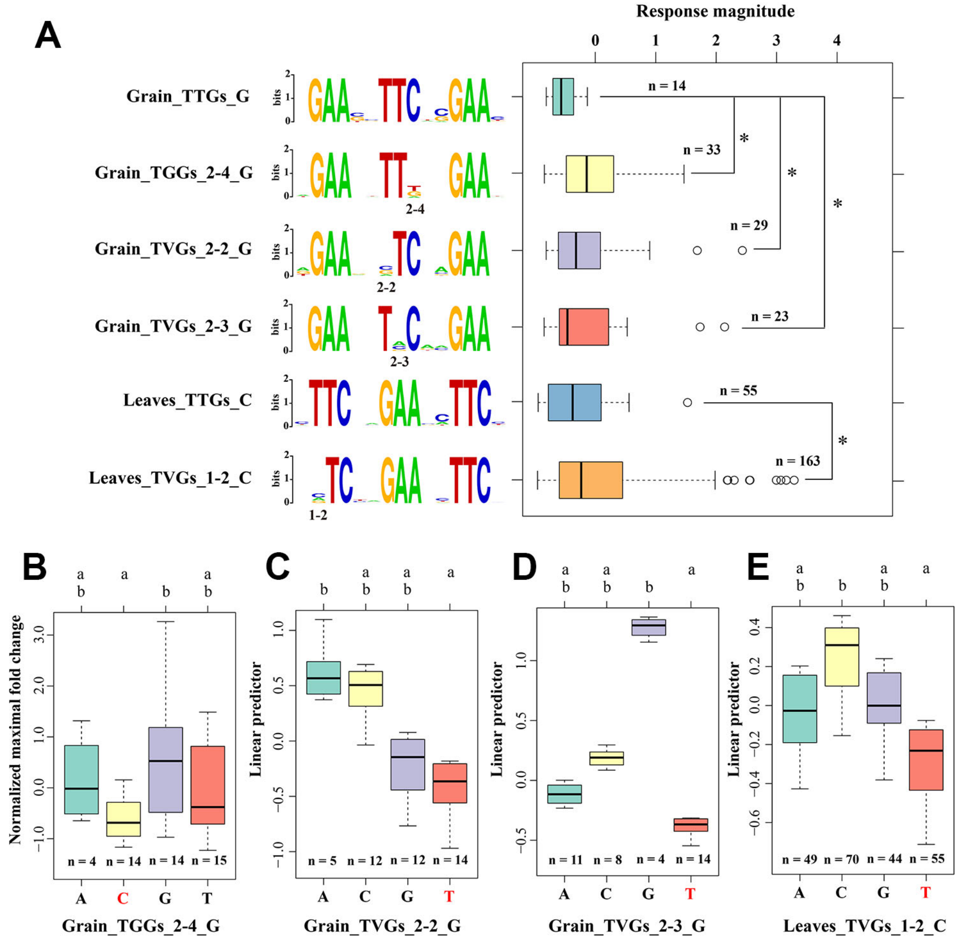


FIGURE 3 | Preference for the position and nucleotide in sequence variations of HSEs **(A)** The effects of different mismatched nucleotides in HSEs on the MFH of CDGs. Compared to the response magnitude of typical HSEs using one-way analysis of covariance with the position of HSEs within the promoter as covariance. "G" and "C" represent the first subunit sequence with 5′-NGAAN-3′ and 5′-NTTCN-3′, respectively. "2-3:" represents the 3rd position in the second subunit. The sequence logo on the middle part illustrates the sequence of each HSE defined on the left. The response magnitude was represented by the normalized maximum fold change value among five heat stress treatment time point. *represent the significant difference of $p < 0.05$. **(B–E)** The effects of different mismatched nucleotides in HSEs on the MFH of CDGs using multiple comparative analysis. The letters above the boxplot represent the significance level and the data sets with different letters were significantly different ($p < 0.05$). Red bases at the x-axis represent the nucleotides of canonical HSEs at this position. Y-axis represents the linear predictor of MFH.

patterns, the involved pathways, and the responsive magnitude between filling grains and flag leaves were distinctly different (Wang et al., 2019). Here, the effects of HSE varying architecture on the response magnitude and the preference for both the position and nucleotide in HSE sequence variations were also different between these two organisms. Because the HSE DNA sequences were equal between these two organism, we assume that different factors, such as the proteins that interacted with HSFs or the transcription initiation complexes, the status of chromatin, and the energy status, resulted in the different effects observed in this study, thus highlighting the differences of gene networks exploited by grain and leaves in HS response.

It is interesting that the vast majority of TTGs, TGGs, and TVGs were not HS-responsive genes and a number of GO terms were also significantly enriched among these genes, implying that these genes and their involved processes may lose their roles in HS response owing to sequence variation of HSEs in their promoter regions or that they do not respond to HS in our investigated organism and HS treatment time points. More importantly, these results motivated us to consider the factors that interact with HSFs to discriminate between HSEs. In the data analysis, several significantly enriched motifs were observed in the promoter regions of HS-responsive genes and non HS-responsive genes, respectively (**Supplemental Tables 3** and **4**), with the AME program (http://meme-suite.org/tools/ame) from MEME package, using the JASPAR CORE 2018 database as background. Interestingly, some enriched motifs were different between the HS-responsive genes and non HS-responsive genes, providing the possibility that other transcription factors may modulate HSF binding to these promoters and affect the

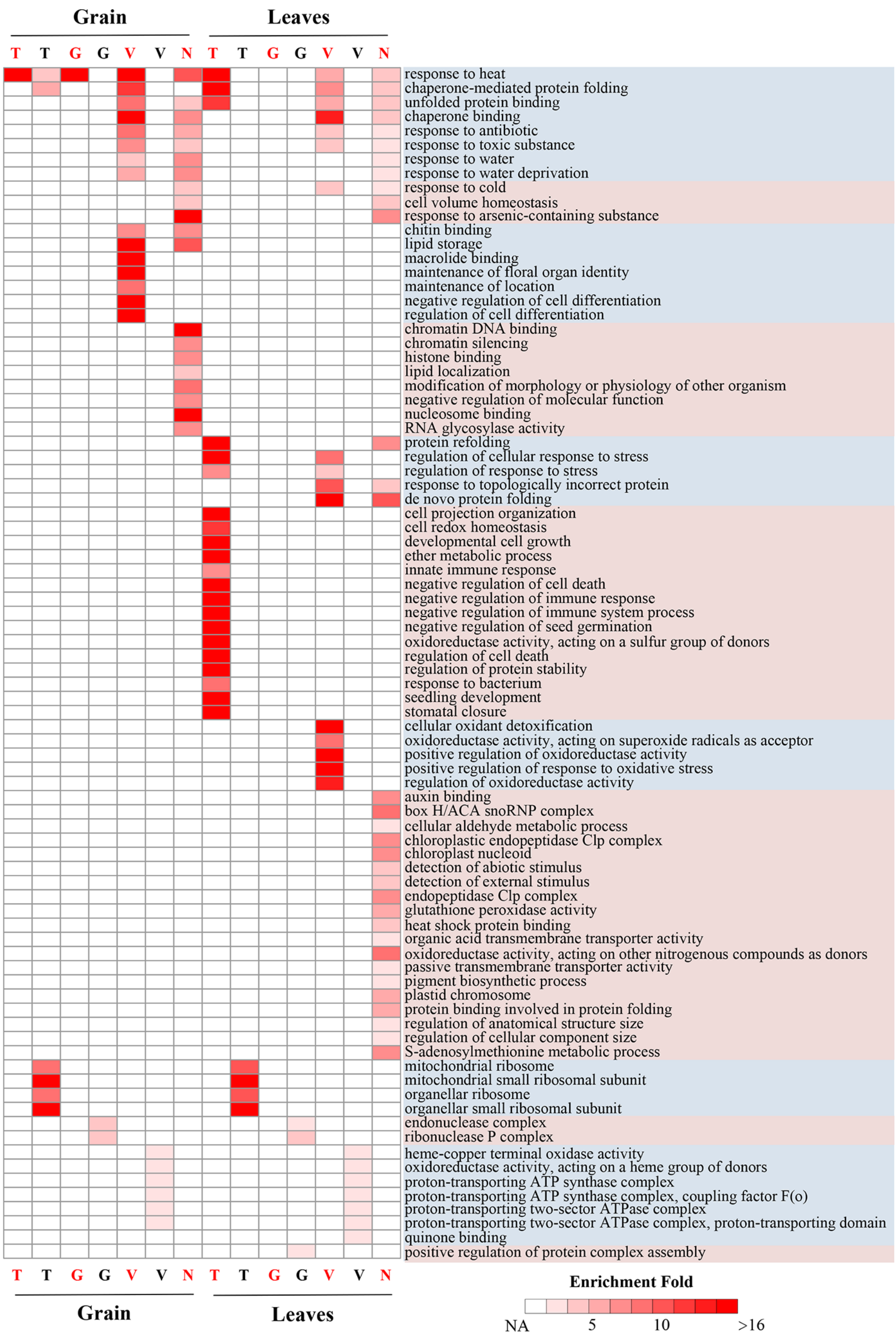


FIGURE 4 | GO enrichment analysis for HS response and non-HS response of TTGs, TGGs, and TVGs GO enrichment analysis for HS response and non-HS response of TTGs, TGGs, and TVGs in grain and leaves. "T," "G," "V," and "N" represent the TTGs, TGGs, TVGs, and genes that do not contain HSEs in their promoter regions, respectively. Red letters at the x-axis represent the HS-response genes. Black letters at the x-axis represent the non-HS response genes. Heat maps show the fold enrichment of enriched GO terms; only significantly enriched terms ($q < 0.01$) are indicated.

subsequent HSF affinity. On the other hand, in *Drosophila*, a ChIP-seq assay showed that HSFs discriminate HSEs based on local signatures of active chromatin (Guertin and Lis, 2010). In wheat, the role of chromatin status variation in HS response was also observed (Liu et al., 2015; Liu et al., 2018), implying that local chromatin status is one of the important factors affecting the interaction between HSF and HSEs. Lastly, other than HS, HSFs were also regarded as core components of signal transduction chains in various abiotic stresses and played a critical role in abiotic stresses response in plants (Guo et al., 2016; Wang et al., 2018b). Therefore, HSEs in the promoter regions of non-HS responsive genes may be discriminated and bound by HSF under other abiotic stresses.

A high efficiency promoter or DNA cis-element in HS response is important not only for the thermotolerance improvement of crops by modulation of gene expression but also for molecular biology studies, because it could be used as an inducible and sharply up-regulated promoter in gene transformation. Although we did not find a specific HSE architecture that continuously had a large response magnitude after heat shock, the finding that varying HSE architecture affects this response provides valuable clues for the directed design of promoters in the future.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

PZ, XS, and BW carried out the data collection. PZ, BW, and XW performed the data analyses. SX and XW contributed to the study design. PZ and XW wrote the manuscript. All authors were involved in the revision of the manuscript and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31501380).

ACKNOWLEDGMENTS

We gratefully acknowledge the support from High performance computing platform of Northwest A&F University (HPC of NWAFU).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00030/full#supplementary-material

REFERENCES

Amin, J., Ananthan, J., and Voellmy, R. (1988). Key features of heat shock regulatory elements. *Mol. Cell Biol.* 8 (9), 3761–3769. doi: 10.1128/mcb.8.9.3761

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc Ser. B-Stat. Methodol.* 57 (1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bokszczanin, K. L., Network, S.P.T.I.T., and Consortium, S. F. (2013). Perspectives on deciphering mechanisms underlying plant heat stress response and thermotolerance. *Front. Plant Sci.* 4, 315. doi: 10.3389/fpls.2013.00315

Busch, W., Wunderlich, M., and Schöffl, F. (2005). Identification of novel heat shock factor-dependent genes and biochemical pathways in Arabidopsis thaliana. *Plant J.* 41, 1–14. doi: 10.1111/j.1365-313X.2004.02272.x

Fernandes, M., Xiao, H., and Lis, J. T. (1994). Fine structure analyses of the *Drosophila* and saccharomyces heat shock factor–heat shock element interactions. *Nucleic Acids Res.* 22 (2), 167–173. doi: 10.1093/nar/22.2.167

Fernandes, M., Xiao, H., and Lis, J. T. (1995). Binding of heat shock factor to and transcriptional activation of heat shock genes in *Drosophila*. *Nucleic Acids Res.* 23 (23), 4799–4804. doi: 10.1093/nar/23.23.4799

Guertin, M. J., and Lis, J. T. (2010). Chromatin landscape dictates HSF binding to target DNA elements. *PloS Genet.* 6 (9), e1001114. doi: 10.1371/journal.pgen.1001114

Guo, L., Chen, S., Liu, K., Liu, Y., Ni, L., and Zhang, K. (2008). Isolation of heat shock factor HsfA1a-binding sites in vivo revealed variations of heat shock elements in Arabidopsis thaliana. *Plant Cell Physiol.* 49 (9), 1306–1315. doi: 10.1093/pcp/pcn105

Guo, M., Liu, J. H., Ma, X., Luo, D. X., Gong, Z. H., and Lu, M. H. (2016). The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Front. Plant Sci.* 7, 114. doi: 10.3389/fpls.2016.00114

International Wheat Genome Sequencing Consortium (IWGSC). (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*. 361 (6403). doi: 10.1126/science.aar7191

Jaeger, A. M., Makley, L. N., Gestwicki, J. E., and Thiele, D. J. (2014). Genomic heat shock element sequences drive cooperative human heat shock factor 1 DNA binding and selectivity. *J. Biol. Chem.* 289 (44), 30459–30469. doi: 10.1074/jbc.M114.591578

Kremer, S. B., and Gross, D. S. (2009). SAGA and Rpd3 chromatin modification complexes dynamically regulate heat shock gene structure and expression. *J. Biol. Chem.* 284 (47), 32914–32931. doi: 10.1074/jbc.M109.058610

Lesk, C., Rowhani, P., and Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature* 529 (7584), 84–87. doi: 10.1038/nature16467

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., et al. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum L.). *BMC Plant Biol.* 15, 152. doi: 10.1186/s12870-015-0511-8

Liu, Z., Qin, J., Tian, X., Xu, S., Wang, Y., Li, H., et al. (2018). Global profiling of alternative splicing landscape responsive to drought, heat and their combination in wheat (Triticum asetivum L.). *Plant Biotechnol. J.* 16 (3), 714–726. doi: 10.1111/pbi.12822

McLeay, R. C., and Bailey, T. L. (2010). Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinf.* 11, 165. doi: 10.1186/1471-2105-11-165

Nguyen, A. D., Gotelli, N. J., and Cahan, S. H. (2016). The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. *BMC Evol. Biol.* 16 (1), 15. doi: 10.1186/s12862-015-0573-0

Ou, J., Wolfe, S., Brodsky, M., and Zhu, L. (2018). motifStack for the analysis of transcription factor binding site evolution. *Nat. Methods* 15, 8–9. doi: 10.1038/nmeth.4555

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi: 10.1093/bioinformatics/btq033

Saidi, Y., Finka, A., and Goloubinoff, P. (2011). Heat perception and signalling in plants: a tortuous path to thermotolerance. *New Phytologist.* 190 (3), 556–565. doi: 10.1111/j.1469-8137.2010.03571.x

Santoro, N., Johansson, N., and Thiele, D. J. (1998). Heat shock element architecture is an important determinant in the temperature and transactivation domain requirements for heat shock transcription factor. *Mol. Cell. Biol.* 18 (11), 6340–6352. doi: 10.1128/mcb.18.11.6340

Tack, J., Barkley, A., and Nalley, L. L. (2015). Effect of warming temperatures on US wheat yields. *Proc. Natl. Acad. Sci. U.S.A.* 112 (22), 6931–6936. doi: 10.1073/pnas.1415181112

Takemori, Y., Enoki, Y., Yamamoto, N., Fukai, Y., Adachi, K., and Sakurai, H. (2009). Mutational analysis of human heat-shock transcription factor 1 reveals a regulatory role for oligomerization in DNA-binding specificity. *Biochem. J.* 424 (2), 253–261. doi: 10.1042/bj20090922

Tao, F., Zhang, Z., Zhang, S., and Rötter, R. P. (2015). Heat stress impacts on wheat growth and yield were reduced in the Huang-Huai-Hai Plain of China in the past three decades. *Eur. J. Agron.* 71, 44–52. doi: 10.1016/j.eja.2015.08.003

Tian, S., Haney, R. A., and Feder, M. E. (2010). Phylogeny disambiguates the evolution of heat-shock cis-regulatory elements in *drosophila*. *PloS One* 5 (5), e10669. doi: 10.1371/journal.pone.0010669

Uffenbeck, S. R., and Krebs, J. E. (2006). The role of chromatin structure in regulating stress-induced transcription in saccharomyces cerevisiae. *Biochem. Cell Biol.* 84 (4), 477–489. doi: 10.1139/o06-079

von Koskull-Doring, P., Scharf, K. D., and Nover, L. (2007). The diversity of plant heat stress transcription factors. *Trends Plant Sci.* 12 (10), 452–457. doi: 10.1016/j.tplants.2007.08.014

Vu, L. D., Gevaert, K., and De Smet, I. (2019). Feeling the heat: searching for plant thermosensors. *Trends Plant Sci.* 24 (3), 210–219. doi: 10.1016/j.tplants.2018.11.004

Wang, X., Wang, R., Ma, C., Shi, X., Liu, Z., Wang, Z., et al. (2017). Massive expansion and differential evolution of small heat shock proteins with wheat (Triticum aestivum L.) polyploidization. *Sci. Rep.* 7, 2581. doi: 10.1038/s41598-017-01857-3

Wang, X., Hou, L., Lu, Y., Wu, B., Gong, X., Liu, M., et al. (2018a). Metabolic adaptation of wheat grain contributes to stable filling rate under heat stress. *J. Exp. Bot.* 69 (22), 5531–5545. doi: 10.1093/jxb/ery303

Wang, X., Shi, X., Chen, S., Ma, C., and Xu, S. (2018b). Evolutionary origin, gradual accumulation and functional divergence of heat shock factor gene family with plant evolution. *Front. Plant Sci.* 9, 71. doi: 10.3389/fpls.2018.00071

Wang, X., Chen, S., Shi, X., Liu, D., Zhao, P., Lu, Y., et al. (2019). Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat. *Plant J.* 98 (6), 1015–1032. doi: 10.1111/tpj.14299

Waters, E. R. (2013). The evolution, function, structure, and expression of the plant sHSPs. *J. Exp. Bot.* 64 (2), 391–403. doi: 10.1093/jxb/ers355

Xiao, H., Perisic, O., and Lis, J. T. (1991). Cooperative binding of *drosophila* heat shock factor to arrays of a conserved 5 bp unit. *Cell* 64 (3), 585–593. doi: 10.1016/0092-8674(91)90242-q

Yu, G., Wang, L., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16 (5), 284–287. doi: 10.1089/omi.2011.0118

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci. U.S.A.* 114 (35), 9326–9331. doi: 10.1073/pnas.1701762114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fgene.2019.01317

The Genetics of Differential Gene Expression Related to Fruit Traits in Strawberry (*Fragaria ×ananassa*)

Christopher Barbey[1]*, Max Hogshead[1], Anne E. Schwartz[1], Nadia Mourad[1], Sujeet Verma[2], Seonghee Lee[2], Vance M. Whitaker[2] and Kevin M. Folta[1]

[1] Horticultural Sciences Department, IFAS, University of Florida, Gainesville, FL, United States, [2] Gulf Coast Research and Education Center, IFAS, University of Florida, Wimauma, FL, United States

Octoploid strawberry (*Fragaria ×ananassa*) is a major specialty crop under intense annual selection for traits relating to plant vigor and fruit quality. Most functional validation experiments rely on transgenic or transient gene expression assays in the mature receptacle. These findings are not typically translatable to breeding without identifying a natural genetic source of transcript level variation, and developing reliable markers for selection in octoploids. Expression QTL (eQTL) analysis is a genetic/transcriptomic association approach for identifying sequence variants predicting differential expression. This eQTL study analyzed a wide array of mature receptacle-expressed genes, encompassing the majority of total mature receptacle transcript accumulation and almost all strawberry genes described in the literature. These results identified segregating genetic variants associated with the differential expression of hundreds of strawberry genes, many with known interest to breeders. Several of these eQTL pertain to published genes whose expression levels have been demonstrated to influence mature receptacle phenotypes. Many include key genes of the phenylpropanoid pathway, vitamin C, carotenoid, pectin, and receptacle carbohydrate/sugar metabolism. These subgenome-specific genetic markers may allow breeders to select for desired ranges of target gene expression. These results may also guide basic research efforts and facilitate the identification of causal genes underlying trait QTL.

Keywords: eQTL analysis, pathway analysis, anthocyanins, pectin, transcriptomics, strawberry (*Fragaria ×ananassa* Duch.)

OPEN ACCESS

Edited by:
Ray Ming,
University of Illinois at Urbana-Champaign,
United States

Reviewed by:
Zhongchi Liu,
University of Maryland,
College Park, United States
Meiru Jia,
China Agricultural University (CAU),
China

****Correspondence:***
Christopher Barbey
cbarbey@ufl.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *06 August 2019*
Accepted: *03 December 2019*
Published: *07 February 2020*

Citation:
Barbey C, Hogshead M, Schwartz AE, Mourad N, Verma S, Lee S, Whitaker VM and Folta KM (2020) The Genetics of Differential Gene Expression Related to Fruit Traits in Strawberry (Fragaria ×ananassa). Front. Genet. 10:1317. doi: 10.3389/fgene.2019.01317

INTRODUCTION

Strawberry is a major specialty crop cultivated worldwide for its sweet and flavorful receptacle, which is referred to commonly as a fruit. Strawberry is under intense breeder selection for new cultivars based on diverse traits. These include receptacle color, firmness, sweetness, yield, flowering time, shipping quality, shelf life, nutrition, flavors, aromas, and disease resistance. The genomics era has provided a dense collection of phenotypically important genes that have been experimentally validated *via* transgenic analysis. However, this basic research often stops short of application, as genetic markers associated with traits are not coincidentally developed for use in breeding. Several resource and technology advances have recently converged to enable high-quality octoploid expression quantitative trait loci (eQTL) analysis. These include an octoploid genome reference (Edger et al., 2019), high-density subgenomic genotyping *via* the IStraw35 platform (Verma et al., 2017), and octoploid reference-based transcriptomics assembly.

eQTL analysis relates genotypic and transcriptomic data to identify segregating genomic regions influencing differential gene expression. Identifying eQTL provides major advantages over pure transcriptomic analysis. The results of an eQTL analysis identify the subset of genes whose differential expression is determined by genotype, the extent of that genetic influence, and markers that may be used for selection of desired gene expression ranges. These selectable markers are potentially useful for application where strawberry phenotypes are known to be influenced by transcript abundance. These include genes which have been characterized *via* transgenic overexpression or silencing in the strawberry receptacle. These eQTL markers may be applied to translate transgenic discoveries into breeding tools. In addition, eQTL controlling transcripts of undetermined function in strawberry can support candidate gene evaluation and trait-based QTL cloning. In one example, simple cross-referencing of trait QTL and eQTL markers identified a causal aroma biosynthesis gene in melon (Galpaz et al., 2018). In strawberry, eQTL experiments helped identify the γ-decalactone biosynthesis gene in the octoploid mature receptacle even while limited to incomplete *de novo* and diploid reference-based RNAseq assemblies (Sánchez-Sevilla et al., 2014). Using the recent subgenome-scale octoploid genome for 'Camarosa', 76 mature receptacle-expressed disease resistances genes (R-genes) were identified to be under the control of an eQTL (Barbey et al., 2019).

Most *cis*-eQTL are caused by sequence variants in or near the gene promoter region (Michaelson et al., 2009). As the approximate causal locus of *cis*-QTL is known, the resolution limits of the IStraw35 genotyping array can be measured. The distance from the originating gene locus to the most-correlated subgenomic marker, when sampled across hundreds of *cis*-eQTL, essentially creates a probability distribution for QTL size-resolution in studies under similar conditions. This distribution can be usefully applied to octoploid QTL studies where *a priori* knowledge of the causal variant locus is not known.

In this work, three octoploid strawberry populations were generated from cultivars varying for fruit quality attributes, such as firmness, sweetness, aroma, and flavor compounds (Vance et al., 2011; Whitaker et al., 2011). Mature receptacle transcriptomes from identical developmental stages were generated and compared against genotype. Analyzed transcripts include those with comparatively high accumulation, those representing differentially expressed genes, and a near-complete list of all published octoploid strawberry genes. Data from the octoploid 'Camarosa' strawberry gene expression atlas (Sánchez-Sevilla et al., 2017) were used to profile the expression of these genes throughout the plant. Genetic associations were filtered based on false-discovery rate (FDR) adjusted *p*-value, effect size, minor allele frequency, and other criteria. Collectively, these results specify the major genetics-based expression differences between cultivars, and the selectable genetics predicting them. These findings bridge basic and applied biology and provide a means to convert previous molecular research directly into plant breeding efforts.

MATERIALS AND METHODS

Plant Materials

Three strawberry flavor and aroma populations were created from Florida cultivars and also 'Mara des Bois' which possesses unique receptacle quality and aroma traits (**Figure S1**). These populations were derived from the crosses 'Florida Elyana' × 'Mara de Bois' (population 10.113), 'Mara des Bois' × 'Florida Radiance' (population 13.75), and 'Strawberry Festival' × 'Winter Dawn' (population 13.76). Mature receptacles were harvested fully ripe from the field during winter growing seasons at the Gulf Coast Research and Education Center (GCREC) in Wimauma, Florida. Populations 13.75 and 13.76 were harvested during the winter of 2014. Population 10.133 was sampled on January 20, February 11, February 25, and March 18, 2011 (Chambers, 2013). Harvest days were selected based on dry weather and moderate temperature, both on the day of harvest and for several days preceding harvest.

Genotyping of Octoploid Strawberry Lines

The Affymetrix IStraw35 Axiom® SNP array (Verma et al., 2017) was used to genotype 61 individuals consisting of parents and progeny from crosses of 'Mara de Bois' × 'Florida Elyana', 'Mara des Bois' × 'Florida Radiance', and 'Strawberry Festival' × 'Winter Dawn' (**Figure S1**). Sequence variants belonging to the Poly High Resolution (PHR) and No Minor Homozygote (NMH) marker classes were included for association mapping. Mono High Resolution (MHR), Off-Target Variant (OTV), Call Rate Below Threshold (CRBT), and Other marker quality classes were discarded and not used for mapping. Individual marker calls inconsistent with disomic Mendelian inheritance from parental lines were removed. Genetic relatedness was evaluated using the VanRaden method using GAPIT v2 package (Tang et al., 2016) in R (**Figure S2**).

Transcriptome Analysis

Octoploid mature receptacle transcriptomes from 61 individuals were sequenced *via* Illumina paired-end RNAseq (avg. 65 million 2× 100-bp reads), consisting of parents and progeny from crosses of 'Florida Elyana' × 'Mara de Bois', 'Florida Radiance' × 'Mara des Bois,' and 'Strawberry Festival' × 'Winter Dawn.' Reads were trimmed and mapped to the *Fragaria ×ananassa* octoploid 'Camarosa' annotated genome (Edger et al., 2019) using CLC Genomic Workbench 11 with a mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.8, similarity fraction of 0.8, and 1 maximum hit per read. Reads which mapped equally well to multiple loci were discarded. RNAseq counts were calculated in Transcripts Per Million (TPM). Transcript levels were normalized *via* the Box-Cox transformation algorithm (Box and Cox, 1964) performed in R-Studio (Racine, 2011) prior to genetic correlation. The BLAST2GO pipeline (Conesa et al., 2005) was used to annotate the full 'Camarosa' predicted gene compliment. Raw reads from the strawberry gene expression atlas study (Sánchez-Sevilla et al., 2017) were aligned to the 'Camarosa' genome using identical procedures, with biological replicates averaged and compared for tissue-based expression using ClustVis (Metsalu and Vilo, 2015) with default parameters.

Identification of High-Variance and Highly Expressed Genes

The 2,000 mature receptacle transcripts with the highest coefficient of variation between samples were identified *via* 1-Pearson correlation distance using the heat map clustering algorithm in CLC Genomics Workbench 11 (**Figure S3**). The 2,000 mature receptacle transcripts with the highest total expression were identified by calculating the sum total expression for each 'Camarosa' transcript across all samples (**Figure S4**).

Retrieval of Published Strawberry Gene Sequences

All 607 non-redundant mRNA accessions under the query "*Fragaria ×ananassa*" were retrieved from the public databases collectively housed at NCBI. This list included all transiently modified strawberry genes compiled in a review (Carvalho et al., 2016) as well as other recently characterized genes in strawberry. Of these, 493 accessions contained an annotated coding sequence (CDS). All retrieved sequences not containing a CDS annotation were determined to be misannotated microsatellite sequences and discarded. These extracted coding sequences were compared by BLAST to identify the most identical gene in the octoploid 'Camarosa' reference genome, identifying 380 unique putative orthologs. This figure was somewhat smaller than the query size, as many deposited mRNA sequences represent alleles or splicoforms corresponding to a single orthologous gene in the non-haplotype specific cv. 'Camarosa' genome. Transcriptome data for these corresponding 'Camarosa' genes were used in eQTL analysis.

Genetic Association of Gene Expression (eQTL)

Genome-wide association was performed *via* a mixed linear model approach using the GAPIT v2 package (Tang et al., 2016) in R. The diploid *Fragaria vesca* physical map was used to orient marker positions, as current genetic maps in octoploid do not include a majority of the available IStraw35 markers. eQTL were evaluated for significance based on the presence of multiple co-locating markers of p-value < 0.05 after false discovery rate (FDR) correction for multiple comparisons (Benjamini and Hochberg, 1995). *Cis* vs *trans* eQTL determinations were made by corroborating the known 'Camarosa' gene position with the eQTL marker position in the physical map.

RESULTS

In all, 268 robust *cis* and *trans* eQTL were discovered relating to the mature receptacle expression of 224 octoploid strawberry genes (**Table S1**). *cis*-eQTL were found abundantly across all subgenomes (**Figure 1**). A vast majority of the identified eQTL loci were found proximal to the originating gene locus, within 0.42 Mb median distance on the same homoeologous chromosome (0.053% of the 'Camarosa' octoploid genome length) (**Table S1**). A frequency plot of *cis*-eQTL (N = 213) marker/gene distances is presented in **Figure 2**. A plurality (16%) of *cis*-eQTL gene/marker distances are located within 0.1 Mb of the originating gene locus. Larger gene/marker distances are progressively rarer until reaching a frequency minimum around 1 Mb. Approximately 90% of gene/marker distances are within this interval. Most eQTL display stepwise changes in transcript accumulation corresponding to allelic dosage, with many displaying near-zero transcript expression in one homozygous state (**Table S1**, **File S1**).

Thirty-five eQTL relate to strawberry alleles known to influence fruit traits *via* transgenic analyses (**Table 1**) or which were experimentally described in strawberry literature (**Table 2**). The magnitude of experimental overexpression/silencing for each gene, and its biological effect in a given cultivar, is shown in comparison with eQTL-associated transcript ranges. In several cases these eQTL naturally replicate transcript accumulation levels observed after transgenic manipulation (**Tables 1** and **2**). These include the published strawberry mature receptacle transcription factors *FanMYB10* (Medina-Puche et al., 2014; Kadomura-Ishikawa et al., 2015; Medina-Puche et al., 2015) *FanEOBII* (Medina-Puche et al., 2015), *FanSnRK2.6* (Han et al., 2015), *FanSLC8* (Pillet et al., 2015), and the phenylpropanoid-modulating genes *FanCCR* (Yeh et al., 2014), *FanF'3H* (Miyawaki et al., 2012), *FanGT1* (Griesser et al., 2008), and *FanFra a3* (Muñoz et al., 2010). Boxplots showing mature receptacle transcript ranges stratified by marker genotype (AA, AB, or BB) are provided for all genes, together with ANOVA omnibus *p* values and *post hoc* significances (**File S1**).

Many of the remaining eQTL-associated genes were further investigated due to their collective participation in mature receptacle quality pathways. These include key genes relevant to phenylpropanoid metabolism (**Figure 3**), flavonoid biosynthesis (**Figure 4**), monolignol biosynthesis (**Figure 5**), and pectin metabolism (**Figure 6**). For fruit phenylpropanoid metabolism, eQTL were discovered for *PHENYLALANINE AMMONIA LYASE, 4-COUMARATE CoA LIGASE, CINNAMATE β -D-GLUCOSYLTRANSFERASE, and CHALCONE 2′-O-GLUCOSYLTRANSFERASE* (**Figure 2**). Relating to fruit flavonoid biosynthesis, eQTL were discovered for *FLAVONOID 3′-HYDROXYLASE, ANTHOCYANIN SYNTHASE, ANTHOCYANIN/FLAVONOL-SPECIFIC UDP-GLUCOSYLTRANSFERASE*, and *DIHYDROFLAVONOL 4-REDUCTASE* (two homoeologs) (**Figure 4**). For fruit monolignol biosynthesis, eQTL were discovered for *HYDROXYCINNAMOYL TRANSFERASE, CINNAMOYL CoA REDUCTASE, 4-COUMARATE CoA LIGASE, ALCOHOL DEHYDROGENASE,*

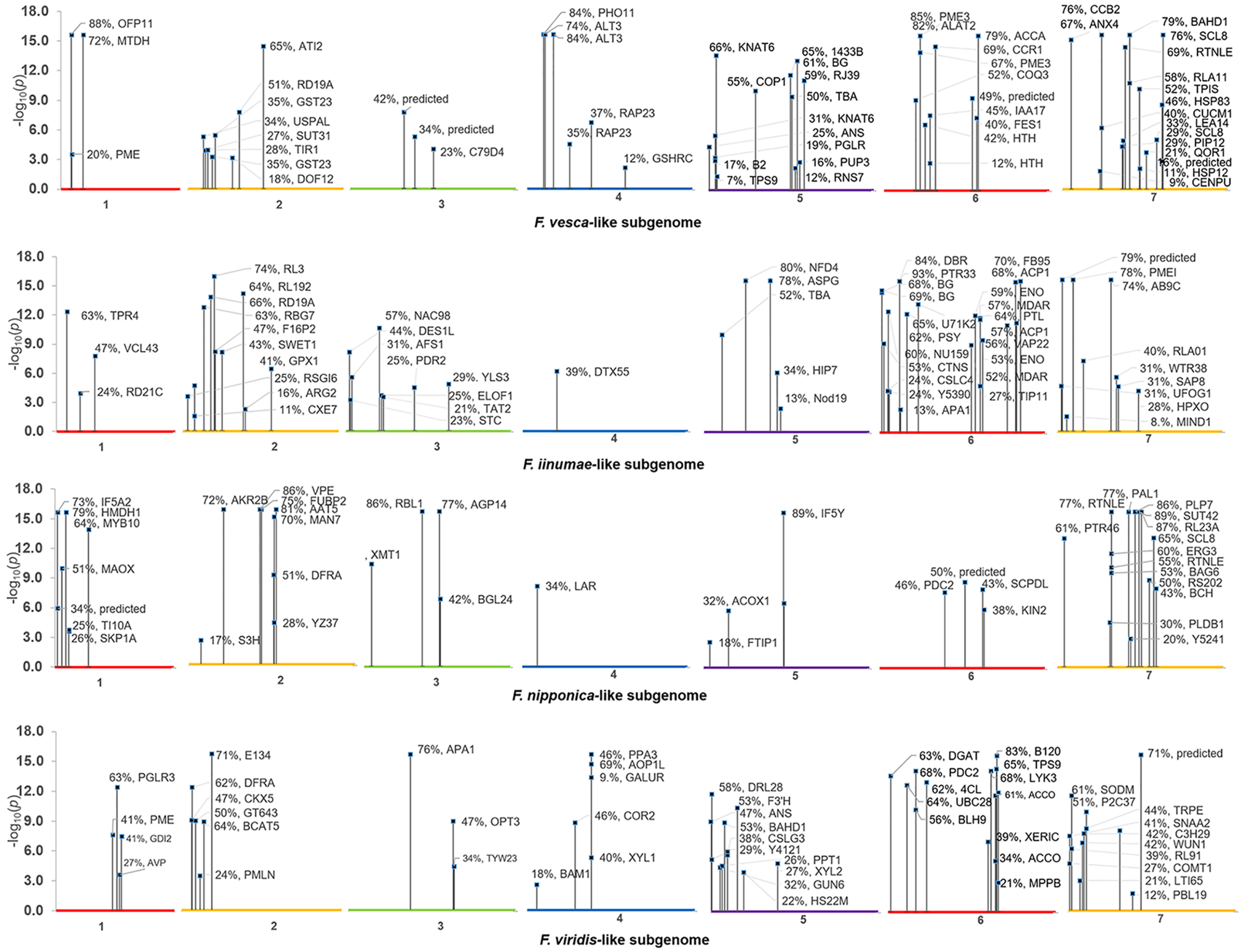


FIGURE 1 | Composite Manhattan plot for octoploid fruit *cis*-eQTL. The 'Camarosa' genome position of the most-correlated marker for each *cis*-eQTL is shown with single-marker *p*-value, effect size and BLAST2GO gene annotation.

and *CAFFEIC ACID O-METHYLTRANSFERASE* (**Figure 5**). These eQTL-associated genes are outlined in context with their pathways, and boxplots demonstrate transcript distribution ranges distributed by marker genotype and with supporting statistics. For fruit pectin metabolism, eQTL were discovered for *PECTIN ESTERASE 3* (two non-homoeologs), *PECTIN METHYLESTERASE INHIBITOR* (two homoeologs and one non-homoeolog), and *POLYGALACTURANASE* (two non-homoeologs) (**Figure 6**). Pectin metabolism-related mature receptacle transcript expression values are shown across parental cultivars (**Figure 6A**) with boxplots of TPM stratified by marker genotype (**Figure 6B**). Genome-wide Manhattan plots for these genes are provided in **File S3**, demonstrating multiple significant markers at each locus.

Other eQTL are highlighted for genes whose transcript levels are known in strawberry to influence sugar/carbohydrate metabolism, **L**-ascorbic acid content, and carotenoid metabolism (**Table 3**). Large mature receptacle transcript abundance differences are dependent upon allele dosage for genes *D-GALACTURONATE REDUCTASE, PHYTOENE CHLOROPLASTIC*, bidirectional sugar transporter SWEET1 and many others. A complete list all 268 eQTL and supporting statistics are presented in **Table S1**, including mean +/- SD transcript values for each marker genotype, minor allele, and minor allele frequencies, transcript variance explained *via* single-marker analysis (omnibus R^2), narrow-sense heritability estimates for mature receptacle transcript accumulation (h^2) with FDR adjusted *p*-values, phase (*cis* or *trans*), physical distance between the originating gene and the *cis*-eQTL, and citations for published genes. Raw transcript abundances in mature receptacles across the eQTL populations (**Table S2**) and in various 'Camarosa' tissues (Sánchez-Sevilla et al., 2017) (**Table S3**) are provided. Complete IStraw35 genotypes for all individuals is provided in **File S2**.

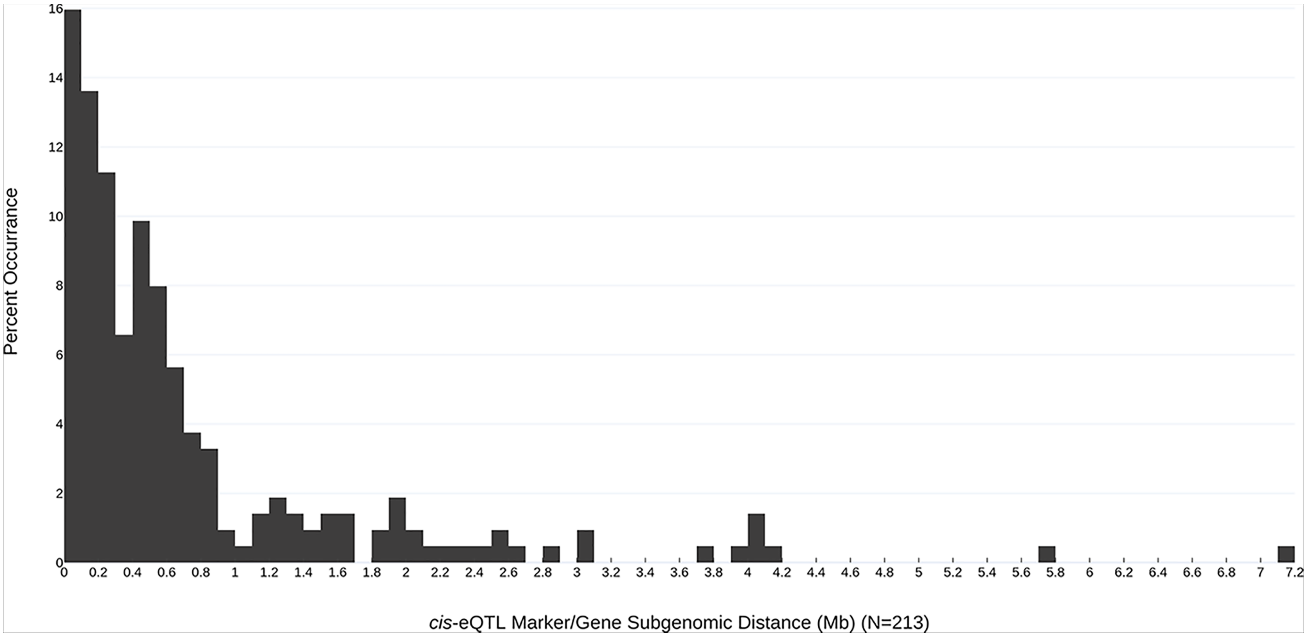


FIGURE 2 | Subgenomic distances (Mb) between the most-correlated *cis*-eQTL marker and the originating gene locus. The frequency of each marker/gene distance observation is indicated (bin size = 0.1 Mb).

Reassembly of raw RNAseq data from various tissues of octoploid 'Camarosa' (Sánchez-Sevilla et al., 2017) determined that a majority of eQTL-associated transcripts predominate in the mature receptacle, and are upregulated with ripening (**Figure 7**). As an external test of the predictive power of each eQTL, the unused 'Camarosa' IStraw35 genotype and six mature receptacle transcriptome replicates (Sánchez-Sevilla et al., 2017) were tested against the eQTL-population transcript distributions for each marker genotype (AA, AB, or BB). Mean 'Camarosa' TPM fell within the 95% prediction interval for its marker genotype in 240 of 268 cases (**Table S1**).

DISCUSSION

This work identified eQTL genetic markers associated with differential mature receptacle transcript accumulation between strawberry genotypes. Most of the identified eQTL are *cis*-variants proximal to the originating gene locus in the 'Camarosa' genome and show stepwise increases in transcript accumulation according to allelic dosage, with many having near-zero TPM in one homozygous state. Loci demonstrating this behavior could correspond to gene presence/absence variants (PAVs) between the cultivars used as parents in this study. Gene PAVs is a major driver of agronomic trait variation in *brassica napus*, with nearly 40% of genes showing PAVs in the pangenome (Hurgobin et al., 2018). Gene PAVs are caused mainly by homeologous exchange (Hurgobin et al., 2018), which has extensively shaped the octoploid strawberry genome (Edger et al., 2019). It seems possible that PAVs could also be a major driver of diversity in octoploid strawberry. Discovery of octoploid eQTL, including those representing PAVs, is limited by the single reference genome available for octoploid transcriptome assembly. This eQTL study hints at the influence of gene PAVs in octoploid cultivar diversity. These eQTL can be immediately leveraged for basic biological investigation, and in some cases genetic selection for several strawberry traits.

Pectin Metabolism

Pectin metabolism is a central feature of ripening-associated sweetening and softening. Pectin metabolism is mainly regulated by differential expression of pectin methylesterases and methylesterase inhibitors (Di Matteo et al., 2005). Several eQTL were identified for these genes, including three pectin methylesterase inhibitors (PMEI), two homoeologs of PECTIN METHYLESTERASE 3 (*PME3*), and two non-homologous POLYGALACTURONASE (*PG*) genes. The eQTL effect size is very strong for the PMEI gene "augustus_masked-Fvb7-3-processed-gene-41.7" (baseline average 29 TPM), where single segregating allele leads to expression increases in excess of 1,500 TPM (>50-fold increase) (R^2 = 0.78, single-marker *p*-value 2.2e-16). This large difference is present among modern cultivars. High PMEI varieties include 'Mara des Bois' and 'Florida Elyana', while low-expression varieties include 'Florida Radiance', 'Winter Dawn', and 'Strawberry Festival'. Transgenic analysis

TABLE 1 | eQTL pertaining to transgenically characterized *F ×ananassa* genes (cv. Camarosa exact putative ortholog).

Gene	Upregulation	Cultivar	Effect caused by overexpression	Cv. Camarosa Gene	eQTL TPM range	Marker (*cis*)
FanCCR (Cinnamoyl-CoA Reductase)	no increased expression	Elsanta and Calypso*	altered content of phenolic acid derivatives	maker-Fvb6-1-augustus-gene-164.26	500-2800	AX-166507442
FanEOBII (Emission of Benzoid II)	2500% to 6500%	Elsanta	increased eugenol content; increased transcript levels of FanEGS2	maker-Fvb6-2-snap-gene-289.57	0-30	AX-166525602
FanFra a3 (Fruit Allergen a 3)	not shown	Elsanta	no effect	augustus_masked-Fvb4-3-processed-gene-199.4	0-8	AX-166513822
FanMYB10 (Myeloblastosis 10)	1.7-fold	Sachinoka	increased anthocyanin content; altered expression of anthocyanin biosynthesis genes	maker-Fvb1-3-augustus-gene-144.30	3-18	AX-123434353
FanSnRK2.6 (Sucrose Nonfermenting1-Related Protein Kinase 2.6)	4-fold	Benihoppe	ripening delay and temperature insensitivity; increased firmness; decreased anthocyanin content; altered transcription of aroma and cell-wall genes.	maker-Fvb2-2-augustus-gene-185.33	0.25-4.5	AX-166521262
Gene	**Silencing (%)**	**Cultivar**	**Effect caused by silencing**	**Cv. Camarosa eQTL Gene**	**eQTL TPM range**	**Marker (*cis*)**
FanCCR (Cinnamoyl-CoA Reductase)	83.3	Elsanta and Calypso*	no effect	maker-Fvb6-1-augustus-gene-164.26	500-2800	AX-166507442
FanEOBII (Emission of Benzoid II)	88.3–99.5	Elsanta	decreased eugenol content; down-regulation of two eugenol-related genes, FanCAD1 and FanEGS2.	maker-Fvb6-2-snap-gene-289.57	0-31	AX-166525602
FanF'3H (Flavonoid 3′-Hydroxylase)	approx. 70	Sachinoka	mild reduction in fruit color	augustus_masked-Fvb5-2-processed-gene-78.0	1-7.5	AX-89893608
FanFra a3 (Fruit Allergen a 3)	60	Elsanta	altered phenylpropanoid pathway precursor and anthocyanin levels; altered transcript levels of FanPAL and FanCHS genes	augustus_masked-Fvb4-3-processed-gene-199.4	0-8	AX-166513822
FanGT1 (Glycosyltransferase 1)	85	Elsanta	mild reduction in color; reduced anthocyanin content; increased flavan-3-ol content	maker-Fvb7-3-augustus-gene-14.53	5-52	AX-166517042
FanMYB10 (Myeloblastosis 10)	90	Elsanta	decreased anthocyanin and eugenol content; decreased transcript accumulation of ripening-related TFs and FanEOBII	maker-Fvb1-3-augustus-gene-144.30	3-18	AX-123434353
	80	Sachinoka	decreased anthocyanin levels; altered transcript levels of flavonoid biosynthesis pathway-related genes			
FanSCL8 (Scarecrow-Like Protein 8)	67 to 93	Strawberry Festival	altered transcript accumulation of flavonoid biosynthesis-related genes	augustus_masked-Fvb7-2-processed-gene-277.8	1-45	AX-166508726
FanSnRK2.6 (Sucrose Nonfermenting1-Related Protein Kinase 2.6)	approx. 90	Benihoppe	acceleration of fruit ripening; increased anthocyanin content; decreased firmness; altered transcript accumulation of pigment, aroma, and cell-wall metabolism genes	maker-Fvb2-2-augustus-gene-185.33	0.25-4.5	AX-166521262

**Transgenic background: FanCHS silenced. Portions of this table are derived from a review by Carvalho et al. (2016).*

can be used to isolate a possible phenotypic effect of this highly variable gene, and determine an ideal allelic state for variety improvement.

Phenylpropanoid Pathway

The phenylpropanoid pathway (PPP) influences many attributes of the strawberry receptacle, and many commercial strawberry breeding priorities are related to phenylpropanoid metabolism. These attributes include firmness and texture, flavor, color, ripening, quantitative disease resistance, shelf-life, and other facets of fruit quality (Dixon et al., 1996; Vogt, 2010; Peled-Zehavi et al., 2015). Several strawberry PPP genes have been characterized using transient expression analysis in the receptacle (Carvalho et al., 2016). Previous RNAseq-based network analyses in strawberry mature receptacles identified that PPP transcripts tend to be highly abundant and broadly variable between cultivars (Pillet et al., 2015). It is expected that PPP-associated transcript levels should vary in populations arising from crosses of strawberry cultivars with contrasting fruit qualities. This eQTL analysis advances previous findings in the strawberry PPP by identifying the specific octoploid subgenomic alleles which are variably expressed due to genetics, and the selectable sequence variants which predict them.

One eQTL in this category is the major transcription factor *FanMYB10* (EU15516, maker-Fvb1-3-augustus-gene-144.30) (R^2 = 0.64, single-marker *p*-value 1.3e-14), which regulates flavonoid and phenylpropanoid metabolism (Medina-Puche et al., 2014). This gene has been studied in strawberry using transgenesis (Kadomura-Ishikawa et al., 2015), but natural variation has not previously been identified. This analysis identifies that the cultivars 'Florida Radiance,' 'Strawberry Festival,' and 'Winterdawn' have 5-to-8-fold greater *FanMYB10* transcript levels compared to 'Mara des Bois' and

TABLE 2 | eQTL pertaining to a published *F ×ananassa* gene (Camarosa' paralog or homoeolog).

Gene	GenBank no.	Cv. Camarosa eQTL Gene	eQTL TPM range	Marker
FanGALUR (D-galacturonate reductase)	AF039182	maker-Fvb4-1-augustus-gene-196.31	500-2600	AX-166505923
FanACP1 (Acyl carrier chloroplastic)	AF041386	maker-Fvb6-3-augustus-gene-389.35 (cis)	1-47	AX-123614270
		(trans)	10-2600	AX-166525307
FanPDC2 (Pyruvate decarboxylase 2)	AF193791	maker-Fvb6-2-augustus-gene-209.38	500-3000	AX-166508206
FanCOBRA (glycosylphosphatidylinositol-anchored protein)	AY642687	maker-Fvb5-2-snap-gene-76.47 (cis)	5-45	AX-166524220
		(trans)	5-45	AX-123361263
FanGT643 (Glycosyltransferase family 64)	AY679583	snap_masked-Fvb2-3-processed-gene-49.22	0.75-2.5	AX-166507157
FanPLDB1 (Phospholipase D beta 1)	AY679584	maker-Fvb7-1-augustus-gene-162.30 (cis)	2-9	AX-166526588
		(trans)	2-9	AX-123365640
FanQOR1 (Quinone oxidoreductase 1)	AY679595	maker-Fvb7-2-augustus-gene-257.57	2-25	AX-166516995
FanYJNA (Uncharacterized AAA domain-containing)	AY679604	maker-Fvb7-4-augustus-gene-17.32	0.25-4	AX-166518372
FanXERIC (RING-type E3 ubiquitin transferase)	AY679613	augustus_masked-Fvb6-4-processed-gene-297.11 (cis)	0.25-7.5	AX-166526717
		(trans)	0.5-7.5	AX-89787062
FanAKR (Aldo/keto reductase)	AY703448	maker-Fvb4-1-augustus-gene-141.33	10000-25000	AX-123367100
FanACCO (ACC oxidase)	AY706156	maker-Fvb6-4-augustus-gene-306.55	70-1400	AX-166516039
FanLEA45 (Late embryogenesis abundant 4-5)	DQ011163	maker-Fvb6-1-augustus-gene-201.55 (cis)	125-1200	AX-166524532
		(trans)	120-900	AX-166518183
FanPPA3 (Purple acid phosphatase 3)	DQ074726	maker-Fvb4-1-snap-gene-183.52 (cis)	2-90	AX-166505923
		(trans)	2-90	AX-166505413
FanF16P2 (D-fructose-1,6-bisphosphate 1-phosphohydrolase)	EU185335	maker-Fvb2-4-snap-gene-100.35	2.5-15	AX-166503535
FanPSY (Phytoene chloroplastic)	FJ784889	maker-Fvb6-3-augustus-gene-80.43 (cis)	10-230	AX-166515961
		(trans)	10-250	AX-89914629
FanZDS (Zeta-carotene desaturase)	FJ795343	maker-Fvb6-2-augustus-gene-256.63 (cis)	5-55	AX-123357007
		(trans)	5-41	AX-166516136
FanNBS1 (TIR-NBS-LRR type protein)	HQ845018	maker-Fvb1-2-augustus-gene-63.28	5-62	AX-89789432
FanNCED1 (9-cis-epoxycarotenoid dioxygenase 1)	JN006161	augustus_masked-Fvb3-3-proces sed-gene-13.7	100-200	AX-166518894
FanRD21C (Probable cysteine protease)	JN979371	maker-Fvb1-2-augustus-gene-106.27 (cis)	100-260	AX-166517617
		(trans)	0-175	AX-89853113
FanMDAR (Monodehydroascorbate reductase)	JQ320104	maker-Fvb6-3-augustus-gene-273.47 (cis)	50-250	AX-123356923
		(trans)	50-150	AX-166513999
FanGR (Glutathione reductase)	JQ339738	maker-Fvb4-3-augustus-gene-315.32	0-0.50	AX-166508304
FanFT1 (FT-like protein)	JQ364958	maker-Fvb6-2-augustus-gene-271.50 (cis)	0-0.50	AX-123363588
		(trans)	200-3900	AX-123364123
FanANS (Anthocyanidin synthase)	JQ923457	maker-Fvb5-1-augustus-gene-7.57	0.50-9	AX-89831030
FanLAR (Leucoanthocyanidin reductase)	JX134096	maker-Fvb4-2-augustus-gene-44.51	25-120	AX-166526717
FanTIR1 (Transport inhibitor response 1)	JX292971	maker-Fvb2-2-augustus-gene-51.44	50-200	AX-89910815
FanG6PDH (Glucose-6-phosphate dehydrogenase cytoplasmic)	KC433888	maker-Fvb6-4-augustus-gene-13.60	18-40	AX-166511756
FanGPX (Glutathione peroxidase 2)	KC433890	maker-Fvb2-4-snap-gene-265.134	18-40	AX-166511756
FanMnSOD (Mn-superoxide dismutase)	KC433893	snap_masked-Fvb7-4-processed-gene-40.42	150-400	AX-123363180
FanDFR (Dihydroflavonol 4-reductase)	KC894054	maker-Fvb2-1-augustus-gene-255.45	50-250	AX-166511816
FanBCH1 (β-carotene hydroxylase 1)	KC967656	maker-Fvb7-1-augustus-gene-290.59	0-1.3	AX-166508808
FanLFY1 (LEAFY-like protein 1)	KF006322	maker-Fvb3-4-augustus-gene-275.43	0-1.8	AX-166513103
FanCERK1 (chitin elicitor receptor kinase 1-like protein)	KT224458	snap_masked-Fvb6-4-processed-gene-308.25 (cis)	1-14	AX-123366408
		(trans)	1-14	AX-123365571
FanMRLK47 (FERONIA-like receptor kinase)	KX374343	augustus_masked-Fvb6-4-processed-gene-67.11	4-29	AX-166508268
FanCOP1 (Constitutive photomorphogenesis 1)	KX583676	maker-Fvb5-1-snap-gene-145.25	1-3.75	AX-123367149
FanPHO11 (Phosphate transporter PHO1 homolog 1)	KY190225	maker-Fvb4-3-snap-gene-46.51	0-1.75	AX-123363868
FanGT2D (trihelix transcription factor)	KY368685	maker-Fvb6-3-augustus-gene-283.340	0-2.2	AX-89868974
FanPIP12 (Plasma membrane intrinsic protein subtype 1 aquaporin)	KY453775	maker-Fvb7-2-augustus-gene-182.44	0-65	AX-123359604
FanJAZ1 (jasmonate ZIM-domain protein)	MF511104	maker-Fvb1-3-augustus-gene-43.42	0-5.5	AX-123365102
FanRJ39 (uncharacterized protein)		maker-Fvb5-1-snap-gene-288.58	125-900	AX-123358407

'Florida Elyana,' and that these differences are heritable (h^2 = 0.93). Relatively modest transient silencing (80-90% of normal) substantially decreased anthocyanin content, whereas relatively modest overexpression (170% of normal) increased anthocyanin content (Kadomura-Ishikawa et al., 2015). The eQTL for *FanMYB10* expression naturally approximates the expression level changes achieved through transgenesis. Genetic selection for this eQTL, and others related to the PPP, could lead to more

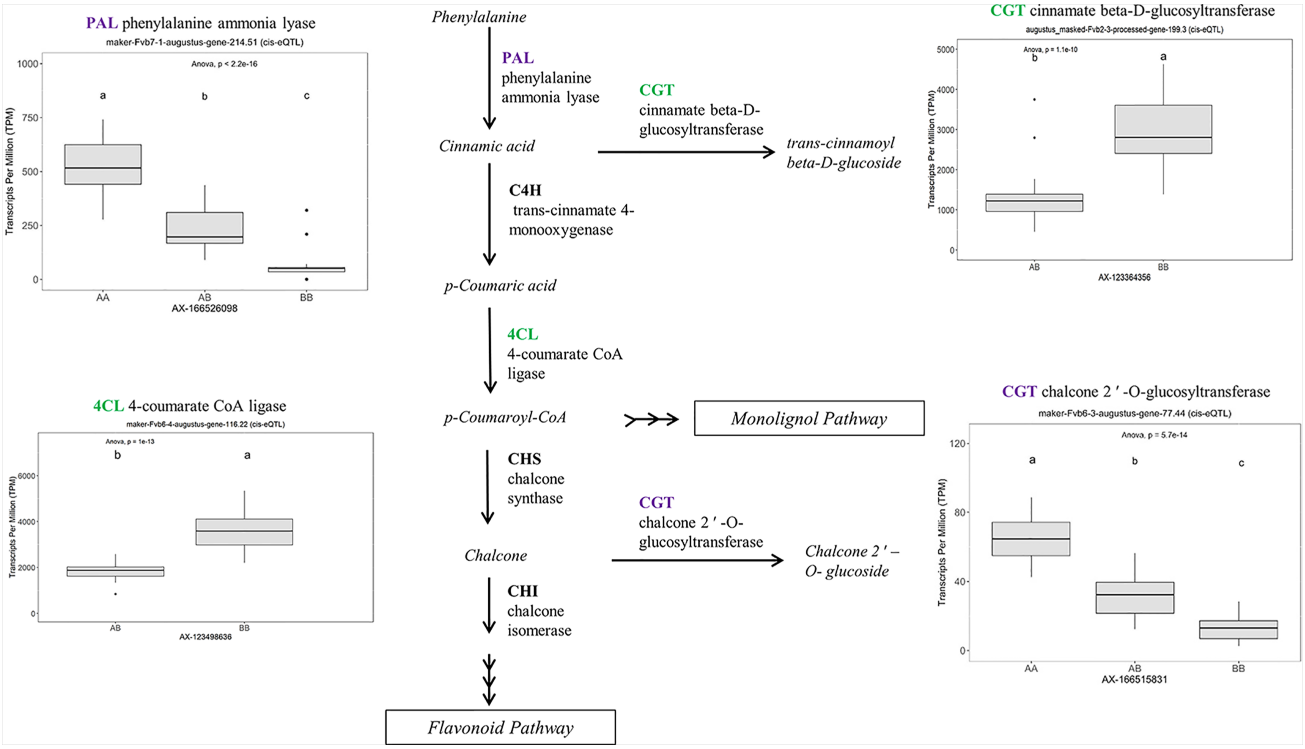


FIGURE 3 | eQTL controlling transcript accumulation of key genes in the strawberry phenylpropanoid pathway (PPP). Marker effect sizes are indicated by boxplots stratified by allelic state (AA, AB, or BB) and shown with ANOVA *p*-values. eQTL genes based on the 'Camarosa' genome are indicated as either possessing the highest sequence identity to the published sequence (purple) or not (green). Letters represent statistically separable means *via* Tukey's HSD post hoc test (p < 0.05).

efficient breeding methods for modified anthocyanin content and other PPP-related metabolites.

A robust eQTL was found for a putative *HYDROXYCINNAMOYL TRANSFERASE* (*FanHCT*, maker-Fvb7-2-augustus-gene-207.46) (R^2 = 0.79, FDR-adjusted *p*-value 0.000023). Hydroxycinnamoyl transferases function in the PPP to generate diverse substrates for Cinnamoyl CoA reductase (CCR) proteins. The candidate *FanHCT* is among the most abundantly accumulating acyltransferases transcripts in the mature receptacle (averaging about 1:200 total transcripts). However, expression of this major transcript is exclusive to 'Camarosa', 'Florida Radiance', and segregating progeny (h^2 = 1.0). Heterologous downregulation of *HCT* expression led to enrichment of H-lignins and improved cell wall saccharification in alfalfa (Jackson et al., 2008), a key process in strawberry receptacle ripening. Several other eQTL were found for other genes in the PPP, including *UDP-GLUCOSE : CINNAMATE GLUCOSYLTRANSFERASE*, an enzyme upstream of HCT.

Vitamin and Nutrient-Associated Transcripts

Both *cis* and *trans* eQTL were identified for *D-GALACTURONIC ACID REDUCTASE* (*FanGalUR*, AF039182; FDR-adjusted p-value 0.0007). The expression of the *FanGalUR* transcript is heritable (h^2 = 0.71) and shows substantial expression variation (500-2,000 TPM range determined by genotype). Previous research in strawberry demonstrated that L-ascorbic acid content is limited by *FanGalUR* transcript abundance (Agius et al., 2003). This limitation has also been confirmed in *F. chiloensis, F. virginiana,* and *F. moschata*, and experimentally validated in species outside of the *Fragaria* genus. Transgenic overexpression of the strawberry *FaGalUR* increased L-ascorbic acid content in *Arabidopsis thaliana* (Agius et al., 2003), *Lactuca sativa L.* (Lim et al., 2008), *Solanum lycopersicum* (Lim et al., 2016), and *Solanum tuberosum* (Hemavathi et al., 2009). It is therefore likely that selecting for increased *FanGalUR* transcript in strawberry will lead to increased L-ascorbic acid content.

L-ascorbic acid levels are influenced by additional factors including metabolite degradation (Cruz-Rus et al., 2011). The gene *MONODEHYDROASCORBATE REDUCTASE* (*MDAR*) is involved in oxidative stress tolerance and is described as a key component of fruit L-ascorbic acid repair (Cruz-Rus et al., 2011). Both *cis* and *trans*-eQTL were discovered for a published strawberry *FanMDAR* allele (JQ320104) (h^2 = 0.68, FDR-adjusted *p*-value 0.00028), of which the *cis*-eQTL accounts for 53% of a 3-fold transcript variation (single-marker *p*-value 8.87e-10). It is possible this heritable fold-change difference contributes to L-ascorbic acid maintenance. This hypothesis could be quickly tested *post hoc* by examining this new genetic variant in

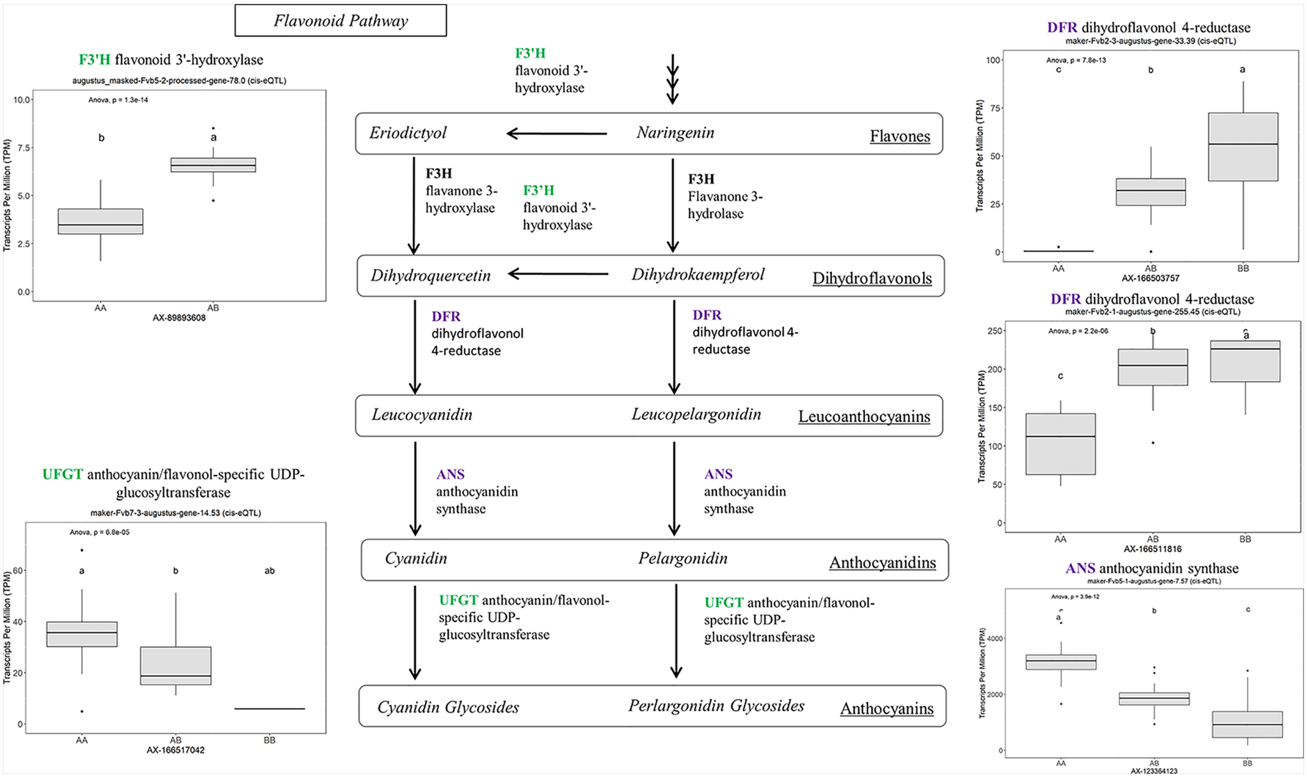


FIGURE 4 | eQTL controlling transcript accumulation of key genes in the flavonoid pathway. Marker effect sizes are indicated by boxplots stratified by allelic state (AA, AB, or BB) and shown with ANOVA *p*-values. eQTL genes based on the 'Camarosa' genome are indicated as either possessing the highest sequence identity to the published sequence (purple) or not (green). Letters represent statistically separable means *via* Tukey's HSD post hoc test (p < 0.05).

strawberry lines with previously existing L-ascorbic acid data. Additional eQTL were found for the vitamin C antioxidant-associated genes *MN-SUPEROXIDE DISMUTASE* (*FanSODM*), *GLUTATHIONE PEROXIDASE* (*FanGPX*), and *GLUTATHIONE REDUCTASE* (*FanGR*) (Erkan et al., 2008) (**Table S2**).

Carotenoids

Strawberry carotenoids provide fruit color and photoprotection, and are essential human nutrients (Ruiz-Sola and Rodríguez-Concepción, 2012). *Cis*-eQTL were discovered for published alleles of strawberry *PHYTOENE SYNTHASE* (*FanPSY*, FJ784889) (h^2 = 0.60, FDR-adjusted *p*-value 0.0018) and *Z-CAROTENE DESATURASE* (*FanZDS*, FJ795343) (h^2 = 0.91, FDR-adjusted *p*-value 0.0034). Single-marker analysis accounts for 62% and 50% of the observed mature receptacle differential expression between genotypes, respectively. *PSY* is a common control point for substrate flux into the carotenoid pathway in several plants (Fraser et al., 2007) and is often correlated with the upregulation of *ZDS* transcripts (Fanciullino et al., 2008). In strawberry fruit, Zhu et al., (2015) noted that *FanPSY* and *FanZDS* transcript accumulation varies between cultivars, and were modestly correlated with carotenoid levels. To assess the impact of *FanPSY* and *FanZDS*-related carotenoid content, these eQTL markers may be used to screen for seedlings that will abundantly express these genes in the fruit. A *cis*-eQTL accounting for 43% of a β-carotene hydroxylase (*FanBCH*) transcript accumulation variance was also discovered, however total accumulation was low (h^2 = 0.92, single-marker *p*-value 1.2e-08).

Fruit Ripening Transcription Factors

The strawberry receptacle ripening process is mainly determined by genetic factors (Perkins-Veazie, 2010). Several eQTL were found for fruit-based transcription factors associated with the ripening process (**Figure 6**), including the negative regulator *FanSnRK2.6,* whose natural expression decreases with fruit ripening (Han et al., 2015). The phenotypic effects of *FanSnRK2.6* expression have been studied *via* transgenesis. Transgenic overexpression of *FanSnRK2.6* (400% of normal) in octoploid receptacles arrested ripening, while silencing (10% of normal) accelerated ripening (Han et al., 2015). The *cis*-QTL for *FanSnRK2.6* is associated with a ~5-fold difference in mature receptacle transcript accumulation between cultivars. As this is a similar range to that demonstrated by transgenesis, it is possible that eQTL marker selection could produce similar phenotypes to that observed by transgenesis. However, the influence of *FanSnRK2.6* is likely greatest in the developing receptacle, and

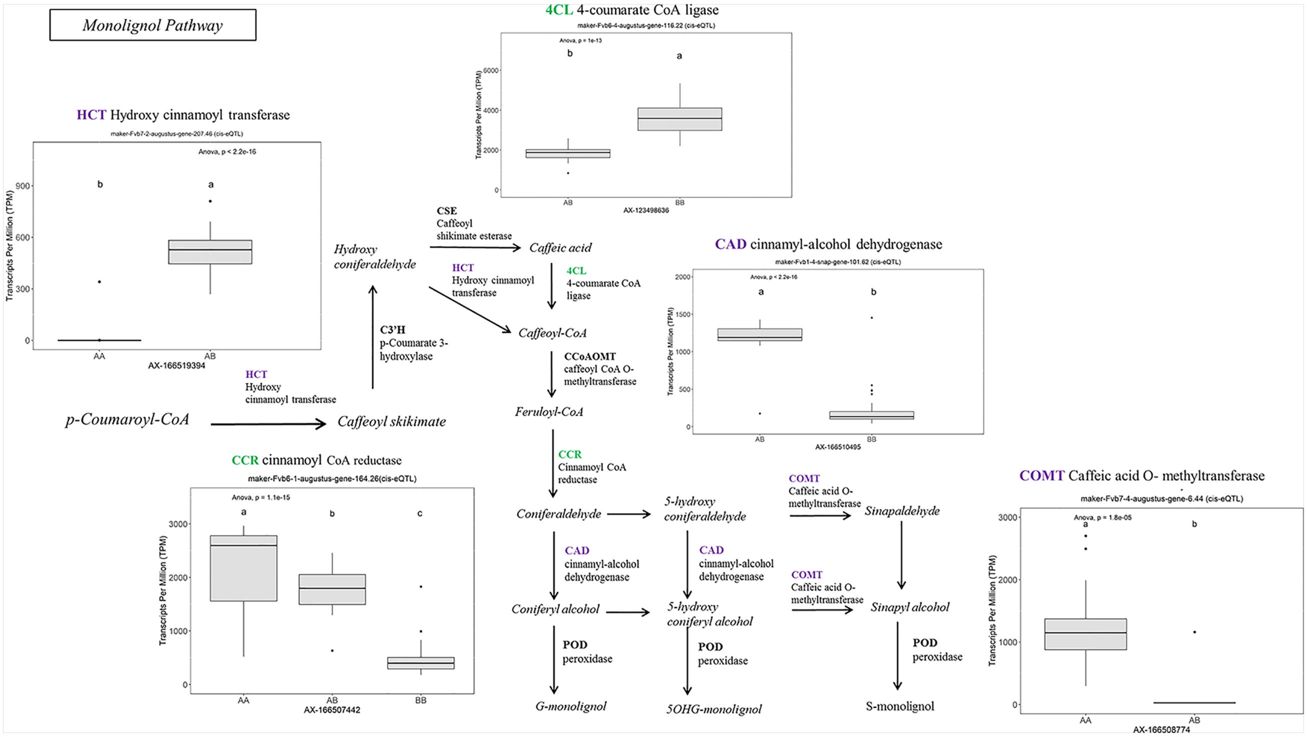


FIGURE 5 | eQTL controlling monolignol pathway gene expression. Marker effect sizes are indicated by box plots stratified by allelic state (AA, AB, or BB) and shown with ANOVA *p*-values. eQTL genes based on the 'Camarosa' genome are indicated as either possessing the highest sequence identity to the published sequence (purple) or not (green). Letters represent statistically separable means *via* Tukey's HSD post hoc test (p < 0.05).

it is unknown if this eQTL is predictive at earlier receptacle stages.

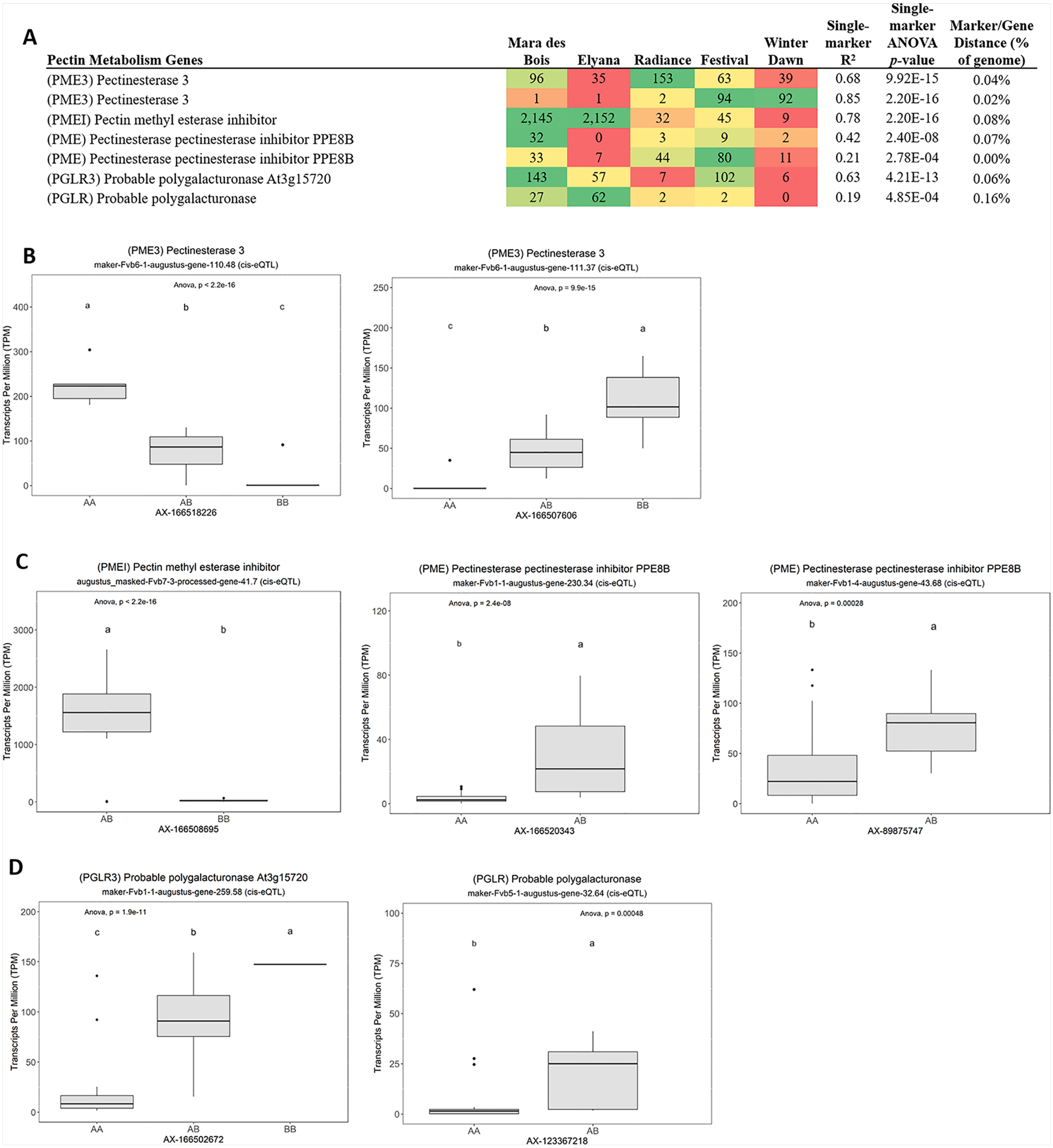
An eQTL was found for the transcription factor *EMISSION OF BENZENOID II* (*FanEOBII, KM099230*). This transcription factor has been experimentally characterized in *Fragaria ×ananassa* using transient overexpression (Medina-Puche et al., 2015). Transient overexpression of *FanEOBII* in the receptacle increased levels of eugenol, a desirable volatile organic compound (Medina-Puche et al., 2015). Ripening-related transcript accumulation of *FanEOBII* is elicited by *FanMYB10*, a phenylpropanoid pathway transcription factor whose eQTL was previously discussed.

An eQTL was also discovered for the flavonoid-associated transcription factor *SCARECROW-LIKE 8* (*FanSCL8, F. vesca-gene13212*). *FaSCL8* was identified as a flavonoid pathway regulator using transcriptome network correlation analysis, and experimentally shown to regulate accumulation of several flavonoid-associated transcripts (Pillet et al., 2015). Additional eQTL were found for one orthologous and one paralogous copy of *FanSCL8*. These gene copies have different expression patterns (Table S1), suggesting non-redundant functions.

Bridging to Application in Strawberry Breeding

eQTL analysis is a tool for evaluating gene expression using genetics. These concrete genetic differences can be used to explore gene function, and serve as a bridge to marker/trait association. The biochemistry and genetics underlying many important traits in strawberry has been detailed in the literature, though few of these discoveries have been translated into practical markers for breeding. It is frequently the case that informative molecular research does not describe a source of beneficial genetics which can be selected by breeders. This eQTL analysis identified natural genetic variants influencing transcript variation, analogous to transgenic expression levels. These markers may also be used for targeted basic research aimed at genes/genetics/transcripts which are highly variable between cultivars. As these eQTL markers are derived from the widely used IStraw35 SNP array platform, eQTL markers can be easily cross-referenced with trait QTL experiments *in silico*. This approach can be used to help identify the causal basis of trait QTL in cases where differential expression contributes to traits. This approach can also be used *post-hoc* to rapidly test existing trait QTL.

These eQTL results highlight the shortcomings of transcriptomics-only driven candidate gene discovery. With RNAseq data alone, it is typically indiscernible whether differential expression is due to genetics, environment or stochastic effects. This genetic association study establishes that most *Fragaria ×ananassa* fruit transcripts are probably not influenced by differential genetics under normal growth conditions. Even among the biased set of 2,000 fruit genes with the highest transcriptional variance, only 8% were rigorously



Pectin Metabolism Genes	Mara des Bois	Elyana	Radiance	Festival	Winter Dawn	Single-marker R^2	Single-marker ANOVA *p*-value	Marker/Gene Distance (% of genome)
(PME3) Pectinesterase 3	96	35	153	63	39	0.68	9.92E-15	0.04%
(PME3) Pectinesterase 3	1	1	2	94	92	0.85	2.20E-16	0.02%
(PMEI) Pectin methyl esterase inhibitor	2,145	2,152	32	45	9	0.78	2.20E-16	0.08%
(PME) Pectinesterase pectinesterase inhibitor PPE8B	32	0	3	9	2	0.42	2.40E-08	0.07%
(PME) Pectinesterase pectinesterase inhibitor PPE8B	33	7	44	80	11	0.21	2.78E-04	0.00%
(PGLR3) Probable polygalacturonase At3g15720	143	57	7	102	6	0.63	4.21E-13	0.06%
(PGLR) Probable polygalacturonase	27	62	2	2	0	0.19	4.85E-04	0.16%

A
B
(PME3) Pectinesterase 3
maker-Fvb6-1-augustus-gene-110.48 (cis-eQTL)
Anova, p < 2.2e-16
a
b
c
Transcripts Per Million (TPM)
0
100
200
300
400
AA
AB
BB
AX-166518226
(PME3) Pectinesterase 3
maker-Fvb6-1-augustus-gene-111.37 (cis-eQTL)
Anova, p = 9.9e-15
c
b
a
Transcripts Per Million (TPM)
0
50
100
150
200
250
AA
AB
BB
AX-166507606
C
(PMEI) Pectin methyl esterase inhibitor
augustus_masked-Fvb7-3-processed-gene-41.7 (cis-eQTL)
Anova, p < 2.2e-16
a
b
Transcripts Per Million (TPM)
0
1000
2000
3000
AB
BB
AX-166508695
(PME) Pectinesterase pectinesterase inhibitor PPE8B
maker-Fvb1-1-augustus-gene-230.34 (cis-eQTL)
Anova, p = 2.4e-08
b
a
Transcripts Per Million (TPM)
0
40
80
120
AA
AB
AX-166520343
(PME) Pectinesterase pectinesterase inhibitor PPE8B
maker-Fvb1-4-augustus-gene-43.68 (cis-eQTL)
Anova, p = 0.00028
b
a
Transcripts Per Million (TPM)
0
50
100
150
200
AA
AB
AX-89875747
D
(PGLR3) Probable polygalacturonase At3g15720
maker-Fvb1-1-augustus-gene-259.58 (cis-eQTL)
Anova, p = 1.9e-11
c
b
a
Transcripts Per Million (TPM)
0
50
100
150
200
AA
AB
BB
AX-166502672
(PGLR) Probable polygalacturonase
maker-Fvb5-1-augustus-gene-32.64 (cis-eQTL)
Anova, p = 0.00048
b
a
Transcripts Per Million (TPM)
0
25
50
75
100
AA
AB
AX-123367218

FIGURE 6 | eQTL pertaining to strawberry pectin metabolism. **(A)** Transcript accumulation in parental lines, GWAS-derived FDR *p* values and narrow-sense heritability estimates, single-marker R^2 and *p* values are shown. Phenotype distributions based on allelic state are shown for **(B)** pectin esterases, **(C)** pectin esterase inhibitors, and **(D)** polygalacturonases. Letters represent statistically separable means *via* Tukey's HSD post hoc test ($p < 0.05$).

TABLE 3 | eQTL pertaining to fruit quality pathway genes.

Fruit Quality Genes	TPM (AA genotype)	TPM (AB genotype)	TPM (BB genotype)	Minor Allele	Minor Allele Frequency	*p*-value (FDR adjusted)	Estimated h^2	Single marker R^2	Single-marker *p*-value
L-Ascorbic Acid Metabolism									
(GALUR) D-galacturonate reductase	597.5 ± 98.1	1218.4 ± 602.6	1969.4 ± 323.4	A	0.29	6.9E-04	0.71	0.10	2.2E-03
(MDAR) Monodehydroascorbate reductase	135.6 ± 37.4	78.9 ± 33.9	18.4 ± 40.2	B	0.24	2.8E-04	0.68	0.53	8.9E-10
Carotenoid Metabolism									
(PSY) Phytoene chloroplastic	36.9 ± 39.0	129.3 ± 36.2	222.9 ± NA	B	0.33	1.8E-03	0.60	0.63	5.5E-13
(ZDS) Zeta-carotene chloroplastic chromoplastic	28.4 ± 12.4	16.4 ± 5.9	7.7 ± 4.2	A	0.42	3.5E-03	0.91	0.50	2.0E-09
(BCH) Beta-carotene 3-chloroplastic	–	.6 ± .4	.1 ± .1	A	0.30	4.02E-02	0.92	0.43	1.16E-08
Monolignol Pathway									
(BAHD1) BAHD acyltransferase At5g47980	7.8 ± 50.3	587.5 ± 250.8	–	B	0.48	2.3E-05	1	0.79	2.2E-16
(CCR1) Cinnamoyl- reductase 1	2864.5 ± 1483.0	1755.4 ± 425.8	455.3 ± 294.3	A	0.23	2.5E-02	0.51	0.69	2.6E-15
(MTDH) Probable mannitol dehydrogenase	–	1150 ± 325.9	197.2 ± 217.3	B	0.26	8.4E-04	0.37	0.72	2.2E-16
Sugar/Carbohydrate Metabolism									
(ENO) Enolase	–	124.8 ± 72.3	6.6 ± 21.6	A	0.29	3.8E-04	0.73	0.59	7.7E-13
(E134) Endo-1,3 1,4-beta-D-glucanase	20.1 ± 12.6	75.4 ± 27.3	167.3 ± 13.8	B	0.36	4.1E-04	0.95	0.71	3.7E-16
(PGLR3) Probable polygalacturonase At3g15720	18.5 ± 30.3	90.5 ± 33.3	182.6 ± 49.9	B	0.31	5.2E-04	0.18	0.63	4.2E-13
(MAN7) Mannan endo-1,4-beta-mannosidase 7	582.7 ± 170.1	322.6 ± 95.2	103.6 ± 100.3	A	0.46	5.8E-04	0.49	0.70	1.2E-15
(AGP14) Arabinogalactan peptide 14	–	479.6 ± 139.0	78.5 ± 78.8	A	0.43	8.0E-04	0.99	0.78	2.2E-16
(SWET1) Bidirectional sugar transporter SWEET1	9.0 ± 35.0	91.4 ± 61.2	–	B	0.19	8.1E-04	0.82	0.43	1.2E-08
(XYL1) Alpha-xylosidase 1	45.5 ± 39.5	4.5 ± 6.8	.5 ± NA	A	0.31	8.3E-04	0.82	0.40	5.1E-06
(XYL2) Beta-xylosidase alpha-L-arabinofuranosidase 2	.2 ± .6	52.6 ± 49.5	–	B	0.33	2.4E-03	0.53	0.28	1.6E-05
(MTDH) Probable mannitol dehydrogenase	–	1150 ± 325.9	197.2 ± 217.3	A	0.20	8.4E-04	0.37	0.72	2.2E-16
(BGL24) Beta-glucosidase 24	52.5 ± 23.2	21.9 ± 24.0	1.9 ± 1.5	A	0.45	1.3E-03	0.76	0.42	1.5E-07
(GUN6) Endoglucanase 6	230.7 ± 88.6	711.9 ± 436.2	651.1 ± 237.6	B	0.40	2.2E-03	1	0.32	4.1E-05
(GBA2) Non-lysosomal glucosylceramidase	–	42.6 ± 26.3	11.2 ± 5.5	A	0.17	4.3E-03	0.29	0.48	9.4E-10
(XYL1) Alpha-xylosidase 1	45.5 ± 39.5	4.5 ± 6.8	.5 ± NA	A	0.31	8.3E-04	0.82	0.40	5.1E-06
(XYL2) Beta-xylosidase alpha-L-arabinofuranosidase 2	.2 ± .6	52.6 ± 49.5	–	B	0.33	2.4E-03	0.53	0.28	1.6E-05
(GPAT3) Probable glycerol-3-phosphate acyltransferase 3	13.5 ± 14.4	54.5 ± 34.0	–	B	0.28	5.4E-03	0.86	0.38	3.0E-07
(BAM1) Beta-amylase chloroplastic	89 ± 39.7	72.9 ± 25.8	29.7 ± 7.4	B	0.19	3.4E-02	0.77	0.19	2.8E-03
(STC) Sugar carrier protein C	73.1 ± 66.9	55.8 ± 43.5	13.5 ± 12.6	A	0.47	4.0E-02	0.75	0.23	6.5E-04
(CSLG3) Cellulose synthase G3	.6 ± 14.6	44.5 ± 20.6	76.7 ± 23.4	A	0.32	4.9E-02	0.82	0.38	9.9E-07
(CSLC4) Cellulose synthase C4	–	165 ± 51.5	100.4 ± 48.8	A	0.25	6.6E-03	0.58	0.25	5.1E-05
trans	–	165.0 ± 51.5	100.4 ± 48.8	A	0.25	6.6E-03		0.25	5.1E-05
(DGAT) Diacylglycerol O-acyltransferase	.5 ± .6	559.5 ± 252.9	–	B	0.35	1.3E-05	1	0.64	2.3E-14
(GLYK) D-glycerate 3-kinase, chloroplastic	97.5 ± 40.1	53.0 ± 11.2	9.8 ± 1.7	B	0.31	9.6E-03	0.45	0.54	1.9E-10
	84.3 ± 39.5	24.1 ± 20.6	–	B	0.12	9.6E-03		0.34	1.0E-06
(PDC2) Pyruvate decarboxylase 2	71.1 ± 33.3	1019.2 ± 527.0	2012.6 ± 510.1	A	0.48	6.9E-04	0.97	0.69	7.2E-15

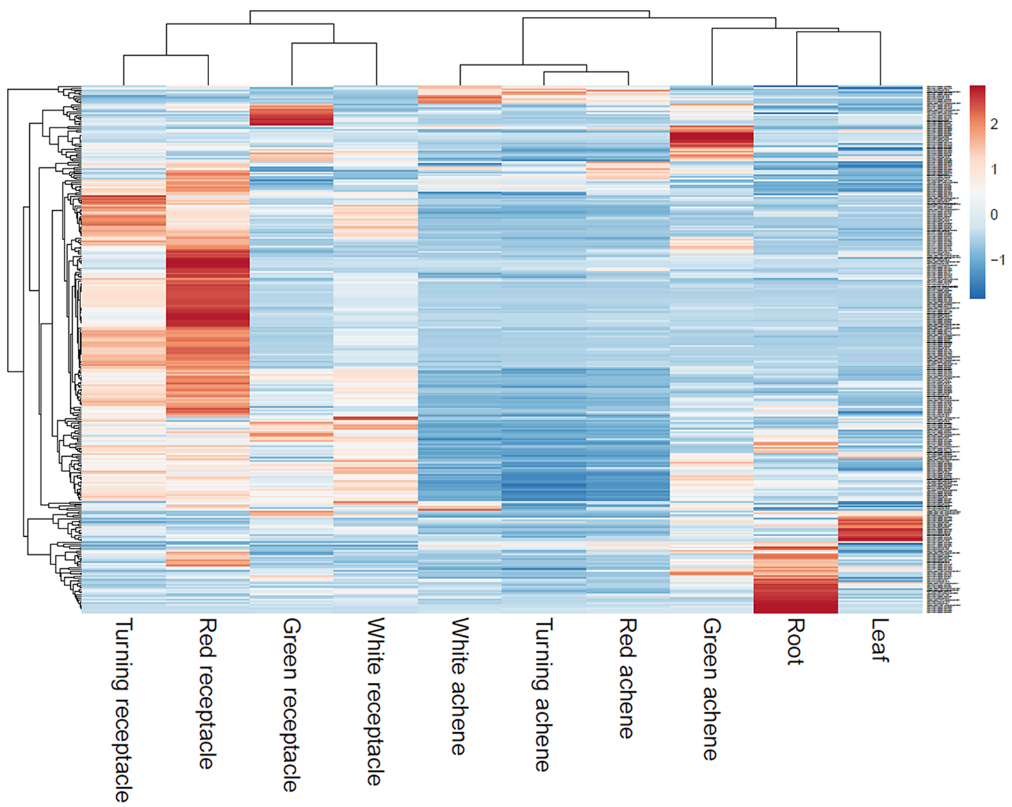


FIGURE 7 | Transcript accumulation of fruit eQTL genes across various tissues. Scaled heatmap of gene expression reveals fruit eQTL genes are predominantly or exclusively expressed in the later fruit stages.

associated with segregating genetics (**Table S1**). As myriad genetic and environmental interactions can influence transcript accumulation, further work is warranted to widen the scope of this foundational analysis. Though multiple populations from different seasons were used in this analysis, these results pertain only to mature strawberry receptacle in normal field conditions, and the modest number of transcriptomes (61) is a limitation for estimating heritability and R^2. Future RNAseq experiments performed in octoploid strawberry are encouraged to utilize low-cost IStraw-based genotyping to facilitate expression-QTL analysis.

AUTHOR'S NOTE

eQTL analysis in octoploid strawberry uncovered genetic variants determining the differential expression of key fruit genes, including published genes where transcript-level variation is known to govern important traits.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Raw short read RNAseq data from fruit transcriptomes are available from the NCBI Short Read Archive under project SRP039356 (http://www.ncbi.nlm.nih.gov/sra/?term=SRP039356). Raw short read RNAseq data from the 'Camarosa' gene expression atlas (Sánchez-Sevilla et al., 2017) are available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) with the study reference PRJEB12420.

AUTHOR CONTRIBUTIONS

CB conceived and led the research experiment. MH, NM, and CB performed gDNA isolation and genotyping data filtering. AS, MH, NM, and CB evaluated eQTL candidates. AS, MH, and CB performed results collection, organization, and single-marker analysis. SV provided guidance in genotyping and QTL mapping. KF, SL, and VW contributed to project oversight and manuscript editing. CB and KF composed the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from the Florida Department of Agriculture and Consumer Services and the UF/IFAS Plant Breeding Graduate Initiative (VW and KF).

ACKNOWLEDGMENTS

The authors acknowledge Drs. Jeremy Pillet and Alan Chambers for RNA isolation and RNAseq line selection, Ben Harrison and Aristotle Koukoulidis for assistance with gDNA isolation, Drs. Iraida Amaya and José Sánchez-Sevilla for raw RNAseq data allowing for expression profiling, and Drs. Pat Edger and Steven Knapp for access to an advanced draft of the octoploid 'Camarosa' genome.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01317/full#supplementary-material

REFERENCES

Agius, F., González-Lamothe, R., Caballero, J. L., Muñoz-Blanco, J., Botella, M. A., and Valpuesta, V. (2003). Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nat. Biotechnol.* 21, 177. doi: 10.1038/nbt777

Barbey, C., Lee, S., Verma, S., Bird, K. A., Yocca, A. E., Edger, P. P., et al. (2019). Disease resistance genetics and genomics in octoploid strawberry. *Genes Genomes Genet.* g3, 400597–402019. doi: 10.1534/g3.119.400597

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc.: Ser. B (Methodol)* 26, 211–243. doi: 10.1111/j.2517-6161.1964.tb00553.x

Carvalho, R. F., Carvalho, S. D., O'Grady, K., and Folta, K. M. (2016). Agroinfiltration of strawberry fruit — a powerful transient expression system for gene validation. *Curr. Plant Biol.* 6, 19–37. doi: 10.1016/j.cpb.2016.09.002

Chambers, A. H. (2013). *Strawberry flavor: from genomics to practical applications*. Dissertation. (Gainesville, FL, USA: University of Florida).

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Cruz-Rus, E., Amaya, I., Sanchez-Sevilla, A., Botella, M. A., and Valpuesta, V. (2011). Regulation of L-ascorbic acid content in strawberry fruits. *J. Exp. Bot.* 62 (12), 4191–4201.

Di Matteo, A., et al. (2005). Structural basis for the interaction between pectin methylesterase and a specific inhibitor protein. *Plant Cell* 17 (3), 849–858.

Dixon, R. A., Lamb, C. J., Masoud, S., Sewalt, V. J. H., and Paiva, N. L. (1996). Metabolic engineering: prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses — a review. *Gene* 179, 61–71. doi: 10.1016/S0378-1119(96)00327-7

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., et al. (2019). Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547. doi: 10.1038/s41588-019-0356-4

Fanciullino, A. L., Dhuique-Mayer, C, Froelicher, Y., Talón, M., Ollitrault, P., and Morillon, R. (2008). Changes in carotenoid content and biosynthetic gene expression in juice sacs of four orange varieties (Citrus sinensis) differing in flesh fruit color. *J. Agric. Food Chem.* 56 (10), 3628–3638.

Fraser, P. D., Enfissi, E. M., Goodfellow, M., Eguchi, T., and Bramley, P. M. (2007). Metabolite profiling of plant carotenoids using the matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Plant J.* 49 (3), 552–564.

Galpaz, N., Gonda, I., Shem-Tov, D., Barad, O., Tzuri, G., Lev, S., et al. (2018). Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping. *Plant J.* 94, 169–191. doi: 10.1111/tpj.13838

Griesser, M., Hoffmann, T., Bellido, M. L., Rosati, C., Fink, B., Kurtzer, R., et al. (2008). Redirection of flavonoid biosynthesis through the down-regulation of an anthocyanidin glucosyltransferase in ripening strawberry fruit. *Plant Physiol.* 146, 1528. doi: 10.1093/jxb/ern117

Han, Y., Dang, R., Li, J., Jiang, J., Zhang, N., Jia, M., et al. (2015). FaSnRK2. 6, an ortholog of Open Stomata 1, is a negative regulator of strawberry fruit development and ripening. *Plant Physiol.* 114, 915–930. doi: 10.1104/pp.114.251314

Hemavathi, U. C. P., Young, K. E., Akula, N., Kim, H. S., Heung, J. J., Oh, O. M., et al. (2009). Over-expression of strawberry d-galacturonic acid reductase in potato leads to accumulation of vitamin C with enhanced abiotic stress tolerance. *Plant Sci.* 177, 659–667 doi: 10.1016/j.plantsci.2009.08.004

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnol. J.* 16, 1265–1274. doi: 10.1111/pbi.12867

Jackson, L. A., Shadle, G. L., Zhou, R., Nakashima, J., Chen, F., and Dixon, R. A. (2008). Improving saccharification efficiency of alfalfa stems through modification of the terminal stages of monolignol biosynthesis. *Bioenergy Res.* 1, 180. doi:10.1007/s12155-008-9020-z

Kadomura-Ishikawa, Y., Miyawaki, K., Takahashi, A., Masuda, T., and Noji, S. (2015). Light and abscisic acid independently regulated FaMYB10 in Fragaria × ananassa fruit. *Planta* 241, 953–965. doi: 10.1007/s00425-014-2228-6

Lim, M.-Y., Cho, Y.-N., Chae, W.-K., Park, Y.-S., Min, B.-W., and Harn, C.-H. (2008). Transgenic lettuce (Lactuca sativa L.) with increased vitamin C levels using GalUR gene. *J. Plant Biotechnol.*

Lim, M. Y., Jeong, B. R., Jung, M., and Harn, C. H. (2016). Transgenic tomato plants expressing strawberry d-galacturonic acid reductase gene display enhanced tolerance to abiotic stresses. *Plant Biotechnol. Rep.* 10, 105–116. doi: 10.1007/s11816-016-0392-9

Medina-Puche, L., Cumplido-Laso, G., Amil-Ruiz, F., Hoffmann, T., Ring, L., Rodríguez-Franco, A., et al. (2014). MYB10 plays a major role in the regulation of flavonoid/phenylpropanoid metabolism during ripening of Fragaria × ananassa fruits. *J. Exp. Bot.* 65, 401–417. doi: 10.1093/jxb/ert377

Medina-Puche, L., Molina-Hidalgo, F. J., Boersma, M., Schuurink, R. C., López-Vidriero, I., Solano, R., et al. (2015). An R2R3-MYB transcription factor regulates eugenol production in ripe strawberry fruit receptacles. *Plant Physiol.* 168, 598. doi: 10.1104/pp.114.252908

Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.* 43, W566–W57010. doi: 10.1093/nar/gkv468

Michaelson, J. J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 48, 265–276. doi: 10.1016/j.ymeth.2009.03.004

Miyawaki, K., Fukuoka, S., Kadomura, Y., Hamaoka, H., Mito, T., Ohuchi, H., et al. (2012). Establishment of a novel system to elucidate the mechanisms underlying light-induced ripening of strawberry fruit with an < i > Agrobacterium-mediated RNAi technique. *Plant Biotechnol.* 29, 271–277. doi: 10.5511/plantbiotechnology.12.0406a

Muñoz, C., Hoffmann, T., Escobar, N. M., Ludemann, F., Botella, M. A., Valpuesta, V., et al. (2010). The Strawberry Fruit Fra a Allergen Functions in Flavonoid Biosynthesis. *Mol. Plant* 3, 113–124. doi: 10.1093/mp/ssp087

Peled-Zehavi, H., Oliva, M., Xie, Q., Tzin, V., Oren-Shamir, M., Aharoni, A., et al. (2015). Metabolic engineering of the phenylpropanoid and its primary, precursor pathway to enhance the flavor of fruits and the aroma of flowers. *Bioeng (Basel Switzerland)* 2, 204–212. doi: 10.3390/bioengineering2040204

Perkins-Veazie, P. (2010). Growth and ripening of strawberry fruit. In Horticultural Reviews, J. Janick (Ed.). doi: 10.1002/9780470650585.ch8

Pillet, J., Yu, H.-W., Chambers, A. H., Whitaker, V. M., and Folta, K. M. (2015). Identification of candidate flavonoid pathway genes using transcriptome correlation network analysis in ripe strawberry (Fragaria × ananassa) fruits. *J. Exp. Bot.* 66, 4455–4467. doi: 10.1093/jxb/erv205

Racine, J. S. (2011). RStudio: A platform-independent IDE for R and sweave. *J. Appl. Econom.* 27, 167–172. doi: 10.1002/jae.1278

Sánchez-Sevilla, J. F., Cruz-Rus, E., Valpuesta, V., Botella, M. A., and Amaya, I. (2014). Deciphering gamma-decalactone biosynthesis in strawberry fruit using a combination of genetic mapping, RNA-Seq and eQTL analyses. *BMC Genomics* 15, 218. doi: 10.1186/1471-2164-15-218

Sánchez-Sevilla, J. F., Vallarino, J. G., Osorio, S., Bombarely, A., Posé, D., Merchante, C., et al. (2017). Gene expression atlas of fruit ripening and transcriptome assembly from RNA-seq data in octoploid strawberry (Fragaria × ananassa). *Sci. Rep.* 7, 13737. doi: 10.1038/s41598-017-14239-6

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9. doi: 10.3835/plantgenome2015.11.0120

Vance, M. W., Tomas, H., Craig, K. C., Anne, P., and Elizabeth, B. (2011). Historical trends in strawberry fruit quality revealed by a trial of University of Florida Cultivars and Advanced Selections. *HortScihorts* 46, 553–557. doi: 10.21273/HORTSCI.46.4.553

Verma, S., Bassil, N. V., van de Weg, E., Harrison, R. J., Monfort, A., Hidalgo, J. M., et al. (2017). *Development and evaluation of the Axiom® IStraw35 384HT array for the allo-octoploid cultivated strawberry Fragaria ×ananassa: International Society for Horticultural Science (ISHS)* (Belgium: Leuven), 75–82.

Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20. doi: 10.1093/mp/ssp106

Whitaker, V. M., Hasing, T., Chandler, C. K., Plotto, A., and Baldwin, E. (2011). Historical trends in strawberry fruit quality revealed by a trial of University of Florida cultivars and advanced selections. *HortScience* 46, 553–557. doi: 10.3233/BR-2011-013

Yeh, S.-Y., Huang, F.-C., Hoffmann, T., Mayershofer, M., and Schwab, W. (2014). FaPOD27 functions in the metabolism of polyphenols in strawberry fruit (Fragaria sp.). *Front. Plant Sci.* 5, 518. doi: 10.3389/fpls.2014.00518

Zhu, H., Chen, M., Wen, Q., and Li, Y. (2015). Isolation and characterization of the carotenoid biosynthetic genes LCYB, LCYE and CHXB from strawberry and their relation to carotenoid accumulation. *Scientia Hortic.* 182, 134–144. doi: 10.1016/j.scienta.2014.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 11 February 2020
doi: 10.3389/fgene.2020.00051

Integrated Analysis of Large-Scale Omics Data Revealed Relationship Between Tissue Specificity and Evolutionary Dynamics of Small RNAs in Maize (*Zea mays*)

Yu Xu[1†], *Ting Zhang*[1,2†], *Yuchen Li*[1] *and Zhenyan Miao*[1,2*]

[1] *State Key Laboratory of Crop Stress Biology for Arid Areas, College of Life Sciences, Northwest A&F University, Yangling, China,* [2] *Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Yangling, China*

Edited by:
Yin Li,
Rutgers, The State University of New Jersey, United States

Reviewed by:
Chong Teng,
Donald Danforth Plant Science Center, United States
Jianhong Xu,
Zhejiang University, China
Paul Fourounjian,
The State University of New Jersey, Newark, United States

***Correspondence:**
Zhenyan Miao
miaozhenyan@163.com

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *16 September 2019*
Accepted: *16 January 2020*
Published: *11 February 2020*

Citation:
Xu Y, Zhang T, Li Y and Miao Z (2020) Integrated Analysis of Large-Scale Omics Data Revealed Relationship Between Tissue Specificity and Evolutionary Dynamics of Small RNAs in Maize (Zea mays). Front. Genet. 11:51. doi: 10.3389/fgene.2020.00051

The evolutionary dynamics and tissue specificity of protein-coding genes are well documented in plants. However, the evolutionary consequences of small RNAs (sRNAs) on tissue-specific functions remain poorly understood. Here, we performed integrated analysis of 195 deeply sequenced sRNA libraries of maize B73, representing more than 10 tissues, and identified a comprehensive list of 419 maize microRNA (miRNA) genes, 271 of which were newly discovered in this study. We further characterized the evolutionary dynamics and tissue specificity of miRNA genes and corresponding miRNA isoforms (isomiRs). Our analysis revealed that tissue specificity of isomiR events tends to be associated with miRNA gene abundance and suggested that the frequencies of isomiR types are affected by the local genomic regions. Moreover, genome duplication (GD) events have dramatic effect on evolutionary dynamics of maize miRNA genes, and the abundance divergence for tissue-specific miRNA genes is associated with GD events. Further study indicated that duplicate miRNA genes with tissue-specific expression patterns, such as *miR2275a*, a phased siRNA (phasiRNA) trigger, contribute to phenotypic traits in maize. Additionally, our study revealed the expression preference of 21- and 24-nt phasiRNAs in relation to tissue specificity. This large-scale sRNAomic study depicted evolutionary implications of tissue-specific maize sRNAs, which coordinate genome duplication, isomiR modification, phenotypic traits and phasiRNAs differentiation.

Keywords: maize, microRNA, phasiRNAs, tissue specificity, genome duplication

INTRODUCTION

Plants are multicellular eukaryotes with diverse tissues comprising various cell types, which carry out common processes essential for survival. However, within the physical context of the tissue environment, cells also exhibit unique functions that help define tissue-specific phenotypes (Edwards and Coruzzi, 1990). These common and tissue-specific processes are ultimately controlled by gene regulatory networks that alter the extent of gene expression. The tissue specificity of these processes is often described based on the expression level of protein-coding genes (PCGs). These analyses only partially capture the variety of processes that distinguish different tissues, due to the ignorance of other regulatory elements, such as small RNAs (sRNAs). Plant sRNAs include microRNAs (miRNAs) and small interfering RNAs (siRNAs), which are typically 20–24 nucleotides (nt) in length. These sRNAs regulate the expression of PCGs involved in a variety of biological processes associated with growth, development, and stress responses in plants (Meyers et al., 2008). MiRNAs also play roles in triggering the biogenesis of secondary siRNAs, termed epigenetically activated siRNAs (easiRNAs), *trans*-acting siRNAs (ta-siRNAs) or phased siRNAs (phasiRNAs) (Allen et al., 2005; Creasey et al., 2014). As miRNAs, phasiRNAs function in a homology-dependent manner to suppress the expression of their targets (Fei et al., 2013). The number of phasiRNA-generating loci (*PHAS* loci) varies substantially among species, ranging from over 800 in wild rice (*Oryza rufipogon*) (Liu et al., 2013b) to less than 30 in *Arabidopsis thaliana* (Columbia-0 ecotype) (Fei et al., 2013). Therefore, the investigation of tissue-specific miRNAs and phasiRNAs will help to understand the regulatory mechanisms of tissue specific phenotypes, thereby facilitating crop improvement via genetic engineering and precision breeding.

Maize (*Zea mays*) is not only one of the most economically important crops but also a model monocot plant for genetic and genomic studies. A growing body of molecular genetics and transgenic studies support the role of sRNAs in regulating diverse agronomic traits in maize (Li et al., 2017; Tang and Chu, 2017). However, the large-scale functional characterization of sRNAs is still considerably lagging in maize compared with the progress made in the model species *Arabidopsis*. In the field of bioinformatics and computational biology, there are still plenty of obstacles to further progress. One of those is the deficiency of comprehensive and high-quality annotation of miRNA genes in maize. To date, only 202 distinct mature miRNAs encoded by 169 miRNA genes in maize have been reported in miRBase (release 22) (Kozomara and Griffiths-Jones, 2014). The number of annotated miRNAs in maize is unexpectedly lower than that in other model plants, such as *Arabidopsis* (349 miRNA genes), soybean (*Glycine max*; 482 miRNA genes), and rice (553 miRNA genes) reported in miRBase. Currently, more than 100 maize sRNA-seq datasets are available in the National Center for Biotechnology Information (NCBI) database. These sRNA-seq datasets are of great value in investigating the expression profiles and potential function of miRNAs. However, these datasets were generated from different laboratories worldwide, which have different sequencing qualities and represent miRNAs in different tissues. Given that the incomplete maize reference genome and impropriate approaches used to identify maize miRNAs ranging from prediction to validation of target cleavage (Jiao et al., 2017; Axtell and Meyers, 2018), it is inevitable that some miRNAs remain to be discovered. Thus, an integrated analysis of sRNA-seq datasets will be very helpful in enriching the set of annotated maize miRNAs, exploring tissue-specific miRNA genes, and uncovering the expression of different miRNA variants.

The complexity of maize genome is another major obstacle. Maize was domesticated from its wild progenitor species, teosinte (*Zea mexiccana*), ~10,000 years ago (Doebley, 2004). It has been proposed that the *Zea* lineage has undergone several rounds of genomic duplication (GDs) events, such as whole genome duplications (WGDs), large-scale segmental duplications (SDs), and tandem duplications (TDs), after its divergence from the lineage that gave rise to sorghum (*Sorghum bicolor*) ~5–12 million years ago (Schnable et al., 2011). GDs are a common phenomenon in plant genome evolution, resulting in the expansion and diversification of many gene families. Post-duplication, many genes have been retained in the genome as paralogous pairs, and individual genes comprising the pair have been either sub-functionalized (partitioning and sharing the original gene function) and/or neo-functionalized (gaining novel functions) via sequence and/ or expression divergence (Freeling, 2009). Studies in different species have revealed that ancient miRNA gene families have evolved via several of GD events (Baldrich et al., 2018). For example, the miR166 family contains seven members in *Arabidopsis*. These miRNA members resulted from WGD, SD, and TD, followed by tissue-specific sub-functionalization, as well as the deletion of one of the members (Maher et al., 2006). Similarly in soybean, post-WGD, the majority of miRNA gene singletons originated from rapid decay, and the retained miRNA gene duplicates evolved slower than singletons, regardless of whether they originated from WGD or TDs (Zhao et al., 2015). With the availability of the maize B73 reference genome sequence and genome resequencing data from dozens of maize accessions, this gramineous crop was made more accessible and has become an important model system suitable for investigation of the evolutionary dynamics and consequences of recurrent GD events.

In this study, we attempted to investigate evolutionary implications of tissue-specific sRNA genes in maize. To achieve this, we broadly collected 195 maize sRNA-seq datasets across 14 tissues and performed an integrated analysis to build a comprehensive list of maize miRNA and phasiRNA genes. A genome-wide study on maize miRNA genes was performed to examine isomiR modification, tissue specificity, and duplication status. We also investigated agronomic traits associated with tissue-specific miRNA genes and selected by artificial improvement. Furthermore, we dissected differentiation of phasiRNAs in relation to tissue specificity.

Abbreviations: DAP, days after pollination; DTA, days to anthesis; DTS, days to silk production; DCL1, DICER-LIKE1; DEMs, differentially expressed microRNAs; DS, drought-stressed; easiRNAs, epigenetically activated small interfering RNAs; GD, genome duplication; GWAS, genome-wide association study; miRNAs, microRNAs; isomiRs, microRNA isoforms; NCBI, National Center for Biotechnology Information; nt, nucleotide; phasiRNAs, phased small interfering RNAs; *PHAS* loci, phasiRNA-generating loci; PCGs, protein-coding genes; SD, segmental duplication; SRA, Sequence Read Archive; SNPs, single nucleotide polymorphisms; sRNAs, small RNAs; siRNAs, small interfering RNAs; TD, tandem duplication; ta-siRNAs, *trans*-acting small interfering RNAs; TPM, transcript per million; TE, transposable element; UIs, unclassified intergenic sequences; WGD, whole genome duplication; WW, well-watered.

MATERIALS AND METHODS

Libraries Construction for High-Throughput Sequencing

Maize B73 inbred lines were grown in pots under a controlled growth chamber (28 °C day / 26 °C night, 14 h light / 10 h dark). Well-watered (WW) seedlings were ensured a normal water supply with 80% of soil field moisture capacity. The drought-stressed (DS) seedlings were subjected to progressive stress by withholding water, and the soil relative water content were maintained at 40% of soil field moisture capacity. The two experimental samples (three biological replicates for each) were collected when three fully expanded leaves appeared, and immediately frozen in liquid nitrogen, and stored at −80 °C for RNA extraction.

Total RNA for small RNA sequencing was extracted using TRIzol reagent (Invitrogen, Carlsbad, USA). Total RNA was separated through 17% denaturing polyacrylamide gels and small RNAs between 10- and 60-nt were collected. Then, 5′ and 3′ RNA adaptors were ligated to small RNAs and followed by reverse transcription to produce cDNAs. These cDNAs were subsequently amplified by PCR and subjected to Illumina sequencing by Novogene Company (http://www.novogene.com/).

Computational Analysis of sRNA-Seq Data and miRNA Annotation

A total of 195 maize sRNA-seq libraries (189 downloaded from NCBI database and six generated in this study) were examined in this study (**Supplementary Table 1**). These libraries were constructed from sRNAs extracted from different tissues, including seedlings, leaves, ears, silks, tassels, anthers, pollen, seeds, embryos, endosperm, stalk, shoot apical meristem, early-prophase meiocytes, and roots at different developmental stages. To detect miRNAs, sRNA-seq datasets were analyzed using a bioinformatics pipeline referring to the recently updated pipeline (Axtell and Meyers, 2018), with slight modifications for the maize genome. Raw miRNA reads were normalized to transcripts per million (TPM) by multiplying a factor of 1,000,000 divided by the total number of mapped reads. For samples with two or more replications, final TPM values were means of all replications in corresponding samples. One mismatch between the genome and sRNA sequence reads was allowed, and structural RNAs (transfer RNAs, ribosomal RNAs, small nuclear RNAs, and small nucleolar RNAs), low abundance RNAs (total abundance ≤100 TPM and abundance in at least one library ≤10 TPM), RNA with irregular sizes (retaining 18–26 nt), and highly repetitive sRNAs (hit number on genome >20) were removed. The modified script of miREAP software (Jeong et al., 2011) was used to predict miRNA precursors and potential pairing of miRNA and miRNA*, and miRNA precursors longer than 300 nt were discarded. For each miRNA precursor, the most abundant miRNA was regarded as canonical miRNA. Variants of canonical miRNAs (isomiRs) were identified using Jasmine pipeline (Zhong et al., 2019) and classified into 5′ isomiR, 3′ isomiR, and polymorphic isomiRs. To distinguish between miRNAs and possible siRNAs, precursor sequences with strong bias (strand bias ≥0.9 and abundance bias ≥0.75) were selected, and stem-loop structures were predicted using CentroidFold (Sato et al., 2009). Strand bias was calculated by dividing the number of reads from the sense strand with the sum of the number of reads from both strands. Abundance bias was calculated as the ratio of miRNA, miRNA* and their isomiRs to the number of reads from the sense strand. Precursor sequences that do not have a typical stem-loop structure (mismatch between the miRNA and miRNA* no more than 5 and asymmetrical bulge number no more than 3) were removed. Identified miRNA precursors were divided into previously annotated precursors and novel precursors by BLAST searching against maize precursors in miRBase v22 (http://www.mirbase.org/ftp.shtml). Maize PCGs and transposable elements (TEs) were downloaded from Gramene database (http://ensembl.gramene.org/Zea_mays).

Identification of Differentially Expressed miRNAs

Raw miRNA reads were normalized to TPM by multiplying a factor of 1,000,000 divided by the total number of mapped reads. Final TPM values were means of three replications. The significance level of differential expression (*P*-value) was determined by the following equation (Audic and Claverie, 1997):

$$P(x|y) \;=\; (\frac{N_2}{N_1})\frac{(x+y)\,!}{x\,!\,y\,!\,(1+\frac{N_2}{N_1})x+y+1}$$

In this equation, N_1 and N_2 were replaced with 1,000,000; x and y represent TPM values under well-watered and drought-stressed conditions, respectively.

Identification of Tissue-Specific miRNA Genes

112 of 195 sRNA-seq libraries representing 14 specific tissues and non-stressed conditions were used to define tissue-specific miRNA genes. Shannon entropy and *Z* score methods were combined to identify miRNA genes specifically expressed in a single tissue. Shannon entropy is used to measure the concentration ratio of gene expression levels in different samples and *Z* score is used to detect outliers. Considering an expression vector x for a miRNA gene, expression vectors x_1, x_2, ..., x_n for n tissues, and an observation x_i for tissue I, the entropy (H) of miRNA genes was calculated as:

$$H = -\sum_{i=1}^{n} E_i log_2 E_i$$

where, E_i is the relative expression of miRNA x for tissue i and is defined as:

$$E_i = x_i / \sum_{i=1}^{n} x_i \,.$$

modH is calculated after using a one-step Tukey's biweight to improve robustness of the expression data as described previously (Kadota et al., 2006). The *Z* score is calculated as:

$$Z_i = (x_i - \mu)/\sigma$$

where, μ is the average expression of miRNA genes $x_1, x_2, ..., x_n$ in all tissues, and σ is the standard deviation. The quantity of the Z score represents the distance between the raw expression and average expression. For tissue-specific miRNAs, we used Shannon entropy modH <1.8 and the maximum Z score (Zmax) >3.

Identification of *PHAS* Loci

The PHASIS software (minDepth was set as 3) (Kakrana et al., 2017) was used to identify *PHAS* loci from 195 sRNA-seq datasets. Candidate *PHAS* loci were selected using $P < 1 \times 10^{-5}$. The summarization of *PHAS* loci was generated using *phasmerge* component. The triggers were identified using *phastrigs* component. Recent studies have shown that preference for production of easiRNAs versus other secondary siRNAs (which include phasiRNAs) may in part be the result of mono-uridylation of 22-nt miRNAs such as miR170 and miR171a which triggers the production of easiRNAs (Zhai et al., 2013; Tu et al., 2015). The defining characteristic that sets easiRNAs apart from other secondary siRNAs is that they arise from transcriptionally active retrotransposons and function in the RNA-directed DNA methylation pathway (Creasey et al., 2014). Therefore, the *PHAS* loci originated from retrotransposons and triggered by 22-nt miR170 and miR171a were excluded from our analyses.

Chromosomal Location and Duplication Analysis of miRNA Genes

The chromosomal distribution of miRNA genes was visualized using MapChart software (Voorrips, 2002). Syntenic blocks were analyzed with the CoGe Synmap program using ZmB73 gramene v4.36 masked coding sequence with default parameters (https://genomevolution.org/coge/SynMap.pl). The miRNA genes with no more than two mismatches with the genome were considered as conserved. *PHAS* loci with the same miRNA triggers and at least two pairs of similar phasiRNAs were considered as conserved. Conserved sequence pairs belonging to syntenic regions were defined as syntenic duplication. Conserved sequences with less than 200 kb between them were considered as tandem duplicates (Holub, 2001).

Identification of miRNA Genes Potentially Related to Agronomic Traits

Datasets from genome-wide association study (GWAS) of 35 agronomic traits in 10 different populations were downloaded (**Supplementary Table 2**) and analyzed using the compressed mixed linear model with the R package GAPIT (Lipka et al., 2012). For each population, single nucleotide polymorphisms (SNPs) were filtered using linkage disequilibrium-based variant pruner implemented in PLINK (Purcell et al., 2007), with the parameters as "–indep-pairwise 1000 100 0.2 –geno 0.1 –maf 0.05". The parameters "–geno" and "–maf" were used for controlling missing rate and minor allele frequency (MAF), respectively. In addition, the first three principal components ("Q" matrix) were also used to control population structure. All SNPs with $P < 1 \times 10^{-5}$ were considered as trait-related SNPs.

Expression Analysis of the miRNA Using Stem-Loop qRT-PCR

Maize B73 plants were cultivated in a phytotron at 25 °C with 65% relative humidity under a 14-h/10-h light/dark cycle. Tissues at the V5 stage (leaves and roots), R1 stage (tassels, ears, anther, pollen, and silks), and 20 days after pollination (DAP) stage (seeds, embryos, and endosperms) were collected and immediately frozen in liquid nitrogen. Samples from at least five plants were pooled for each biological replicate, and three biological replicates were performed. Total RNA was extracted following above described methods. To confirm the miRNA expression data, stem-loop qRT-PCR was performed referring to previously described methods (Chen et al., 2005). Typically, 200ng of RNA was digested by DNase I (Thermo scientific, USA) and reverse transcribed using a TaKaRa Mir-X miRNA First-Strand Synthesis Kit (Cat. No.638313). The stem-loop RT-PCR was performed on the ABI7500 Real-Time System (Applied Biosystems, USA) using EvaGreen 2 × qPCR MasterMix-ROX (Abm, Canada). The reactions were incubated in a 96-well plate at 55 °C for 2 min, 95 °C for 10 min, followed by 40 cycles of 95 °C for 30 s and 60 °C for 1 min. All reactions were assayed in three biological replicates, and 18S rRNA was used as the internal control for stem-loop RT-PCR. Normalized expression levels were calculated as previously described methods (Schmittgen and Livak, 2008). Primers used in the qRT-PCR are listed in **Supplementary Table 3**.

Identification of miRNA Target

Potential target transcripts of miRNAs were predicted using psRNATarget (2017 update; http://plantgrn.noble.org/psRNATarget) (Dai and Zhao, 2011), psRobot (v1.2; http://omicslab.genetics.ac.cn/psRobot/downloads.php) (Wu et al., 2012), and TargetFinder (https://github.com/carringtonlab/TargetFinder) (Fahlgren and Carrington, 2010) with default parameters, except for the maximum expectation of five in psRNATarget. Predicted sRNA targets were verified with degradome sequencing data of maize B73. The raw degradome sequencing datasets were downloaded from the NCBI's Sequence Read Archive (SRA) database (SRR768486, SRR768488, SRR768490, SRR768493, SRR895786, SRR895787, SRR895788, SRR895789, SRR1028862, SRR1028864, SRR1028865, SRR1028866, SRR2917866, SRR4183498, SRR4183499, SRR3347501, SRR3348073, SRR3470769, and SRR3470770). Raw reads were preprocessed using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit) to remove adaptors, low-quality bases (score <20), and short reads (<18 nt). Clean reads and maize cDNA sequences were input into CleaveLand4 (setting: -p 0.05) (Addo-Quaye et al., 2009) to identify potential cleavage sites.

Statistical Analysis

The Student's t-test was performed using *t.test* function in R package. The χ^2 test was performed using *chisq.test* function in R package.

RESULTS

Genome-Wide Identification of Maize miRNA Genes and Their Mature Sequences

In this study, we developed a bioinformatics pipeline to identify maize miRNAs and their mature sequences referring to the recently updated pipeline (Axtell and Meyers, 2018) (**Figure 1**). This miRNA identification pipeline was applied to process ~3.83 billion sequencing reads from 195 maize sRNA-seq datasets covering diverse tissues and conditions (**Supplementary Table 1**). As a result, we identified a total of 419 maize miRNA genes (**Supplementary Table 4**), consisting of 148 annotated **miRNA** genes in miRBase v22, as well as 271 miRNA genes newly discovered in this study. Like the miRNA genes present in miRBase, these novel miRNA genes mapped throughput the maize B73 reference genome with an uneven distribution on 10 chromosomes, and most of these genes were in regions of chromosome arms (**Figure 2A**). The sequence structures of 419 maize miRNA genes were visualized in Supplementary Data 1. DICER-LIKE1 (DCL1) is one of central components of the plant miRNA biogenesis pathway. By analyzing previously published *dcl1* mutants sequencing data (Petsch et al., 2015), we found the expression levels of both known and novel miRNAs were reduced in mutation samples (**Supplementary Figure 1**). Of the 419 miRNA genes, 125 (29.8%) were located within PCGs, 125 (29.8%) were surrounded by TEs, and 196 (46.8%) were located within unclassified intergenic sequences (UIs; **Supplementary Table 5**).

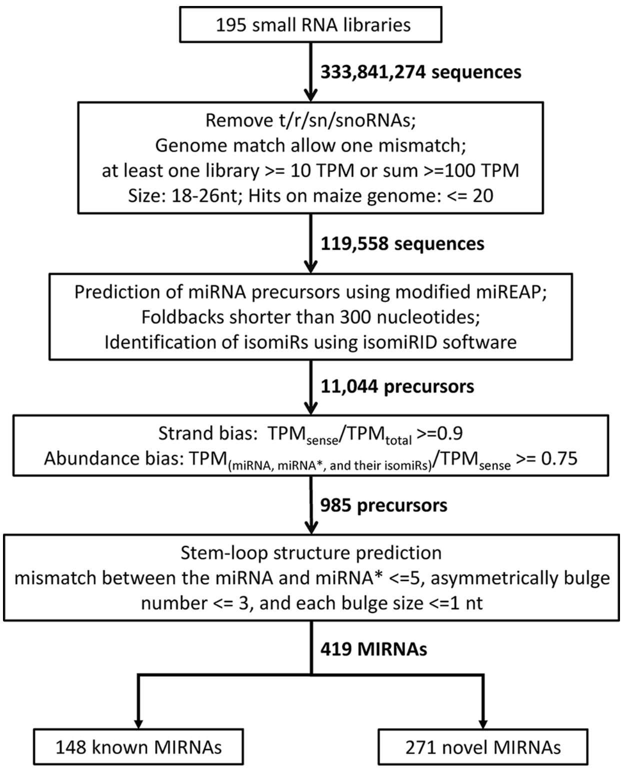

195 small RNA libraries
333,841,274 sequences
Remove t/r/sn/snoRNAs;
Genome match allow one mismatch;
at least one library >= 10 TPM or sum >=100 TPM
Size: 18-26nt; Hits on maize genome: <= 20
119,558 sequences
Prediction of miRNA precursors using modified miREAP;
Foldbacks shorter than 300 nucleotides;
Identification of isomiRs using isomiRID software
11,044 precursors
Strand bias: TPM_{sense}/TPM_{total} >=0.9
Abundance bias: $TPM_{(miRNA,\ miRNA^*,\ and\ their\ isomiRs)}/TPM_{sense}$ >= 0.75
985 precursors
Stem-loop structure prediction
mismatch between the miRNA and miRNA* <=5, asymmetrically bulge number <= 3, and each bulge size <=1 nt
419 MIRNAs
148 known MIRNAs
271 novel MIRNAs

FIGURE 1 | Bioinformatics pipeline for the identification of miRNA genes in maize.

Growing evidence suggested that a single miRNA gene can generate multiple mature isomiRs that differ in their length and/or sequence composition (Zhai et al., 2013). From the 419 miRNA genes, we identified 20,315 isomiRs, which can be categorized into seven classes: nucleotide substitutions within the seed sequence (seed SNP; 41.1%), nucleotide substitutions within tail sequence (tail SNP; 32.8%), 3′ trimming (10.7%), 5′ trimming (8.1%), 5′ addition (7.5%), 3′ addition (6.8%), and non-templated 3′ addition (3′ nt addition; 3.1%) (**Figure 2B**). Biased nucleotide composition was observed at the ends of isomiRs. For instance, the 5′ and 3′ ends of truncated isomiRs were likely to be ‘G' and ‘C' nucleotides, respectively. By Contrast, nucleotides on the 3′ end of non-template added isomiRs were generally ‘A' and ‘U' (**Supplementary Figure 2**).

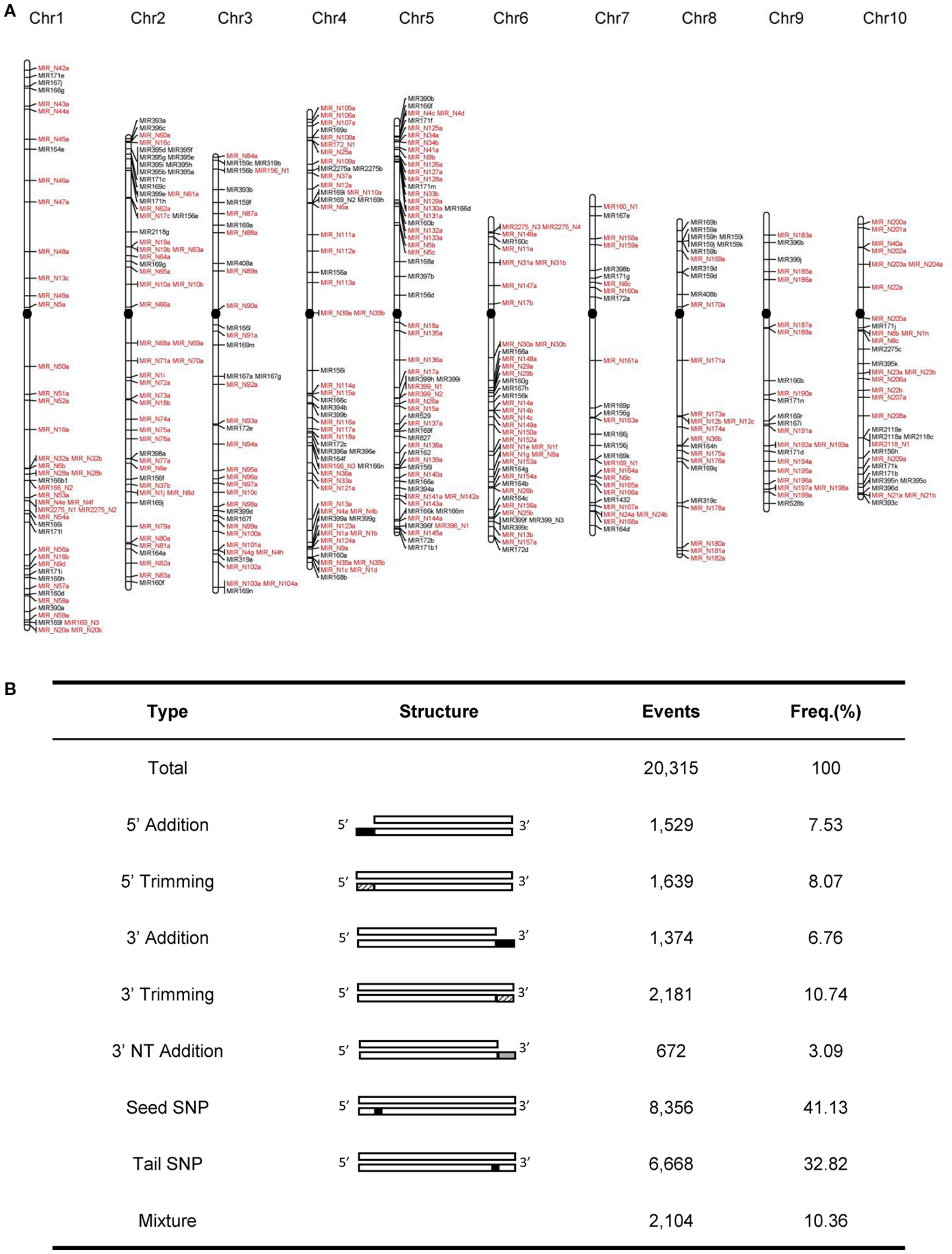
Together, the large-scale analysis of sRNA-seq datasets provides a more comprehensive profiling of maize miRNA genes and the corresponding isomiRs (**Supplementary Table 6**). An overview of the maize sRNA profiling supported by JBrowse (Buels et al., 2016) can be accessed in the Maize sRNA Data Browser, which is publicly available at http://bioinfo.nwafu.edu.cn/MSDB/index.html.

Survey of Differential Expressed miRNAs in Maize Response to Drought Stress

To study the expression dynamics of miRNAs and their potential roles in gene expression regulation in maize drought responses, we constructed and sequenced six sRNA libraries from WW and DS maize B73 inbred lines. Of the 419 miRNAs genes, 43 differentially expressed miRNAs (DEMs) induced by drought stress were detected (**Supplementary Table 7**). To understand the potential regulatory roles of drought stress-associated miRNAs, we identified the target genes of DEMs using three bioinformatics programs applying different alignment algorithms, then further validated by using sequencing datasets from 19 publicly available PARE libraries of maize B73. Of the 43 DEMs, 24 were found to target 30 transcripts (**Supplementary Table 8**). These include 13 transcription factors, such as MYB domain-containing, NAC domain-containing, Scarecrow like, and squamosa promoter binding, as well as 17 other functional genes, such as ubiquitin, zinc ion binding, and Syg1/Pho81/XPRI.

Ninety-Four miRNA Genes Show Highly Tissue-Specific Expression Patterns

A tissue-specific miRNA gene is defined as one that is highly expressed in a specific tissue. Based on the expression profile of maize sRNAs in 14 specific tissues under normal conditions, we identified a total of 94 miRNA genes (34 previously annotated and 60 newly discovered miRNA genes) using Shannon entropy and *Z* score. These genes were specifically expressed in nine tissues (**Figure 3** and **Supplementary Figure 3**). We noticed that endosperm, pollen, and leaf tissues required relatively more specific miRNA genes than other tissues. Some miRNA gene families (e.g. *miR164* and *miR169*) were predominantly



Type	Structure	Events	Freq.(%)
Total		20,315	100
5′ Addition		1,529	7.53
5′ Trimming		1,639	8.07
3′ Addition		1,374	6.76
3′ Trimming		2,181	10.74
3′ NT Addition		672	3.09
Seed SNP		8,356	41.13
Tail SNP		6,668	32.82
Mixture		2,104	10.36

FIGURE 2 | Genome-wide characteristics of maize miRNA genes. **(A)** Chromosomal distribution of 419 maize miRNA genes. Newly identified miRNA genes were labeled as red. The black filled circles represent centromeres **(B)** Statistics of different isomiR types in the maize genome.

expressed in multiple tissues. For example, *miR164e* was specifically expressed in root, whereas *miR164a*, *miR164b*, *miR164c*, *miR164d*, and *miR164g* showed silk-specific expression. Similarly, *miR169m*, *miR169n*, and *miR169q* were preferentially expressed in the leaf, whereas *miR169a*, *miR169b*, and *miR169r* were preferentially expressed in the silk. Then, we performed qRT-PCR assay to further confirm the expression levels of nine miRNAs. As shown in **Supplementary Figure 4**, the results of qRT-PCR assay are consistent with those of sRNA-seq. These results indicated that tissue-specific miRNA gene family members, which differ by only a few nucleotides may have redundant or diverse functions.

Dynamic Patterns of Isomir Events in Terms of Tissue Types and Genomic Environments

Considering the tissue specificity of miRNA expression and regulation, we are interested that whether isomiRs exhibit distinctive properties in different tissues. Statistical analysis showed that the abundance of isomiRs varied widely among tissues (**Figure 4A**). There were more isomiRs events in the ear and stalk tissues, and the abundance of isomiRs varied during development of leaf tissue (**Figure 4A**). The composition of seven isomiR types changed remarkably among the individual tissues (**Supplementary Figure 5**). Nucleotide bias was also observed at the 5′ and 3′ ends of isomiRs among tissues (**Supplementary Figure 6**). Correlation analysis using data from all individual tissue samples showed that the abundance of isomiRs were highly correlated with the abundance of expressed miRNA genes located in PCGs, TEs, and UIs (**Figures 4B–D**). Comparisons of isomiR abundance among

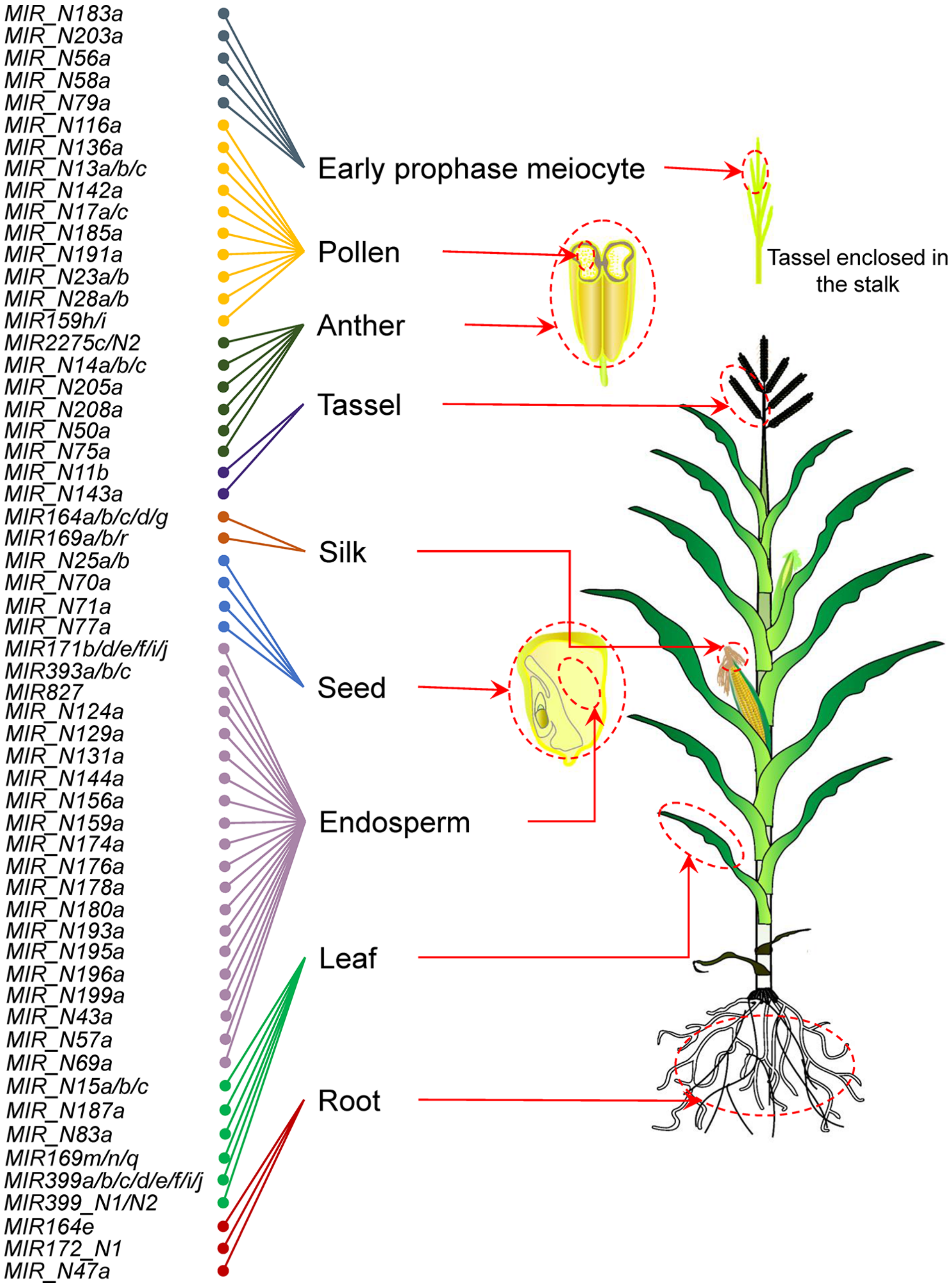


FIGURE 3 | Overview of tissue-specific patterns of maize miRNA genes.

seven isomiR types were conducted in three categories of miRNA genes (PCGs, TEs, and UIs) separately. For miRNA genes located in PCGs, 5′ and 3′ isomiR types (5′ addition, 5′ trimming, 3′ addition and 3′ trimming) exhibited significantly higher abundance than polymorphic isomiR types (seed SNP and tail SNP; **Figure 4E**). In contrast, the abundance of 5′ and 3′ isomiR types were significantly lower for miRNA genes harbored by TEs and UIs (**Figures 4F, G**). These observations indicated that tissue specificity of isomiR events tends to be associated with miRNA gene abundance and suggested that the frequencies of isomiR types are affected by the local genomic regions.

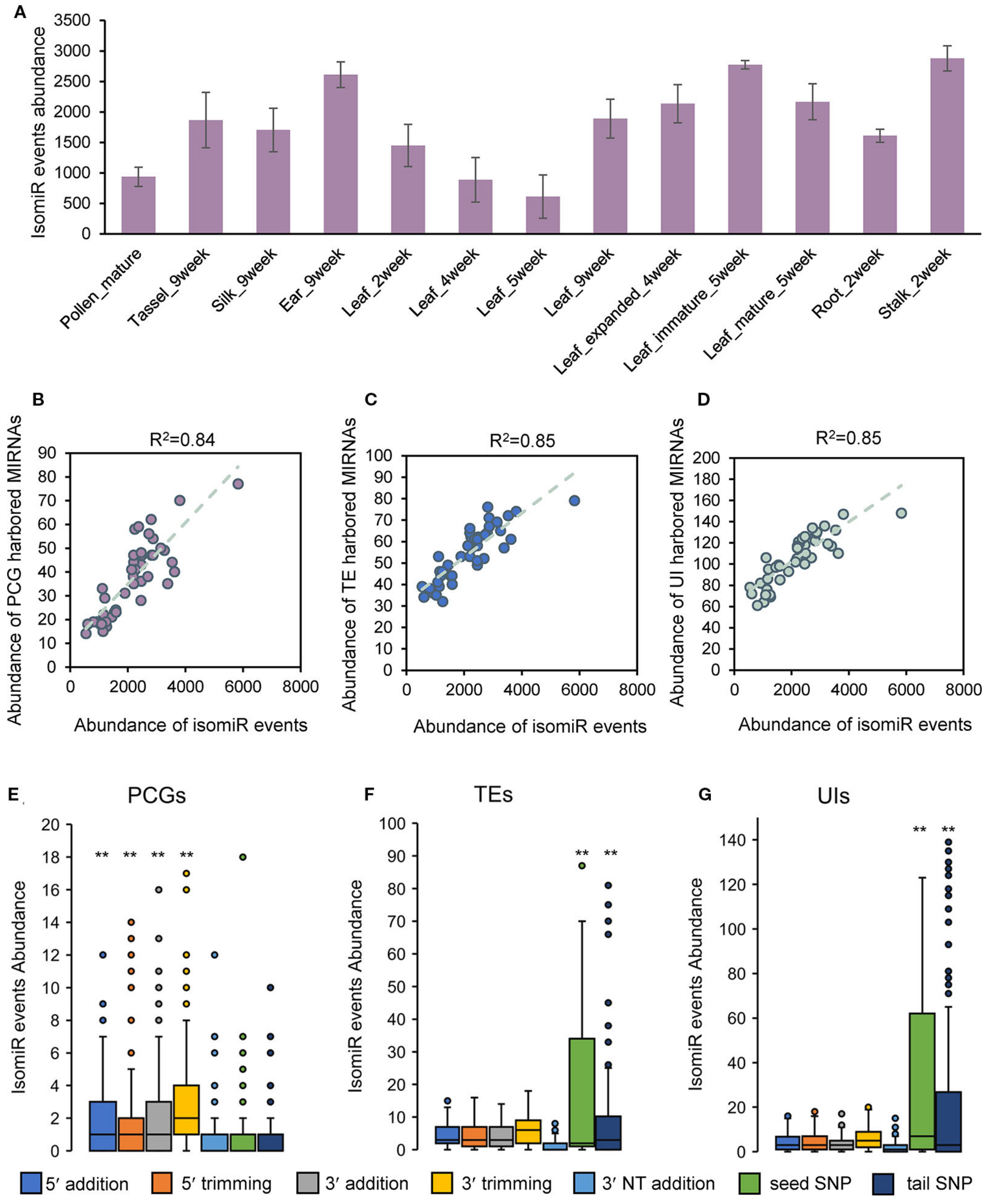


FIGURE 4 | Dynamic patterns of isomiR events in terms of tissue types and genomic environments. **(A)** Abundance of isomiRs varied widely among individual tissues. **(B)** Correlation between the abundance of isomiR events and PCG-harbored miRNA genes in all tissue samples. **(C)** Correlation between the abundance of isomiR events and TE-harbored miRNA genes in all tissue samples. **(D)** Correlation between the abundance of isomiR events and UI-harbored miRNA genes in all tissue samples. **(E)** Comparisons of isomiR abundance among seven isomiR types in PCG-harbored miRNA genes. **(F)** Comparisons of isomiR abundance among seven isomiR types in TE-harbored miRNA genes. **(G)** Comparisons of isomiR abundance among seven isomiR types in UI-harbored miRNA genes. Statistically significant differences in **(E, F** and **G)** were determined using Student's *t*-test and **indicated *P-value* < 0.05.

Evolutionary Implications of Duplicated miRNA Genes on Tissue Development

It has been proposed that maize underwent a series of GD events after its divergence from sorghum ~5–12 million years ago (Swigonova et al., 2004; Schnable et al., 2011). GDs are recognized as essential sources of gene family expansion and functional specialization (Panchy et al., 2016). Due to evolutionary dynamics of duplicated genes, one copy of many duplicated gene pairs in maize was eliminated, leaving the other one as a singleton after the genomic duplication events. Therefore, the singleton-to-duplicate ratio represents the degree of fragmentation of duplicated genomes. To understand the evolutionary process and consequences of tissue-specific miRNA genes after GD, based on the newly annotated set of PCGs in B73 reference genome (Jiao et al., 2017), we reanalyzed the large GDs encompassing the 419 miRNA genes and found that these genes can be classified into 305 singletons and 114 duplicates, with a ratio of 2.68:1 (**Table 1**). For miRNAs located in PCGs, TEs, and UIs, the corresponding singleton-to-duplicate ratio were 8.62:1, 2.79:1, and 1.76:1, respectively (**Table 1**). The chi-square "goodness-of-fit" test showed that the singleton-to-duplicate ratios for miRNA genes in PCGs and UIs were significantly different from that for miRNA genes in the whole maize genome (**Table 1**). In addition, the average abundance of miRNA genes in PCGs was the lowest compared with those in TEs and UIs, and the genome-wide miRNA duplicates showed relatively higher expression than singletons (**Supplementary Table 9**). These analyses suggested that the duplication status of miRNA genes was affected by their locations in the genome.

For the 94 tissue-specific miRNA genes, 72 were singletons and 22 were duplicates, corresponding to a singleton-to-duplicate ratio of 3.27:1, which was significantly higher than that in the whole maize genome (2.68:1; χ^2 test, $P < 0.05$). This indicated an overrepresentation of singletons compared with duplicates in tissue-specific miRNA genes. To further investigate if and how GD contributed to the tissue specificity in maize, we compared the relative proportions of miRNA singletons and duplicates expressed in individual tissue samples. The ratios of expressed miRNA singletons to duplicates (ranging from 0.82:1 to 2.28:1) in all tissues, except endosperm, were significantly lower than that in the whole maize genome (χ^2 test, $P < 0.05$) (**Table 2**). With leaf development, the ratio of expressed miRNA singletons to duplicates decreased after 2 weeks after germination and then increased after 6 weeks, however, the opposite trend was observed during silk development. Moreover, the singleton-to-duplicate ratio decreased during ear development, and the similar trend was also observed in the tassel (**Table 2**). We also compared the ratios of expressed miRNA singletons to duplicates located in the three categories of genomic compartments in each individual tissue sample. In general, among the three categories of miRNA genes, the singleton-to-duplicate ratio was the lowest in UIs and the highest in PCGs in most tissues (**Supplementary Table 10**). Collectively, these observations revealed extensive divergence in the duplication status of miRNA genes among tissues.

TABLE 1 | Statistical analysis of singleton-to-duplicate ratios of miRNA genes located in three categories of genomic components.

	Singleton	Duplication	Ratio (S/D)[a]	P^b
Total	305	114	2.68:1	
PCG	112	13	8.62:1	9.68×10^{-5}
TE	92	33	2.79:1	0.858
UI	125	71	1.76:1	0.023

[a] *The ratio of singletons to duplicates.*

[b] *The statistical analysis was conducted on the ratios of singletons to duplicates between total miRNAs and miRNAs in each category of genomic components by χ^2 goodness of fit test.*

Evolution of Tissue-Specific miRNA Genes Contributed to Phenotypic Traits in Maize

Maize was domesticated from its wild relative teosinte approximately 10,000 years ago and was subsequently subjected to intensive breeding efforts to improve its adaptation to modern agricultural practices, resulting in dramatic morphological and physiological modifications (Lyu, 2017). Recent studies demonstrated that SNPs that occurred at miRNA-related regions (pre-miRNAs and mature miRNAs) may affect miRNA biogenesis and function and cause serious phenotype changes (Sun et al., 2009; Hung et al., 2012; Liu et al., 2013a). Therefore, we investigated whether miRNA loci and quantitative trait loci were simultaneously selected by linkage drag, or whether miRNA loci have effects on tissue-specific traits. To address this question, we collected 10 representative maize B73 populations comprising and integrated the results of GWAS on the genetic diversity in various traits, such as ear height, leaf width, days to anthesis (DTA), days to silk production (DTS), and plant height (**Supplementary Table 2**).

We detected 13 trait-associated mutation sites co-localizing with seven miRNA loci (**Supplementary Figure 7** and **Figure 5A**). Of these seven miRNA loci, four loci were associated with at least two traits. An interesting example is *miR2275a*, which belongs to the anther-specific *miR2275* family and is associated with DTA and DTS (**Figure 5A**). According to our previous analyses of duplicated miRNA loci in maize, we observed that *miR2275_N1*/*miR2275_N2*, *miR2275_N3*/*miR2275_N4*, and *miR2275a*/*miR2275b* were three homoeologous pairs formed by TD (**Figure 5B**). These results promoted us try to understand the process of selection of *miR2275* family members during evolution. Divergencies of the dominant arms were observed among the three homoeologous pairs (**Figure 5C**). Further analysis revealed that *miR2275c* co-localized with a helitron transposon, implying that TE-mediated mechanisms were responsible for the amplification of this miRNA gene family. To shed light on how genomic duplication affects the functional diversity of these homoeologous pairs, we analyzed the expression levels of seven *miR2275* loci in individual tissues. All *miR2275* loci were highly abundant in early-prophase meiocytes and anthers (**Figure 5D**). These observations, together with the association analysis, indicated that although these miRNA loci dominated in anthers, the evolutionary consequences of the *miR2275* family may have been impacted by duplication events leading to functional divergence of individual members. Subsequently,

TABLE 2 | Statistical analysis of singleton-to-duplicate ratios of miRNA genes in individual tissue samples.

	Singleton	Duplicate	Ratio (S/D)[a]	*P*[b]
Total	305	114	2.68:1	
Anther	142	78	1.82:1	0.031
Ear_6week	72	64	1.12:1	1.64×10^{-5}
Ear_8week	74	68	1.09:1	5.39×10^{-6}
Ear_9week	72	74	0.97:1	2.17×10^{-7}
Ear_10week	55	54	1.02:1	8.21×10^{-6}
Early prophase meiocytes	131	85	1.54:1	3.91×10^{-3}
Embryo_1d	87	63	1.38:1	7.84×10^{-4}
Embryo_3week	57	46	1.24:1	5.78×10^{-4}
Embryo_9DAP	85	73	1.16:1	1.38×10^{-5}
Embryo_15DAP	89	66	1.35:1	4.24×10^{-4}
Embryo_20DAP	67	61	1.10:1	1.42×10^{-5}
Endosperm_9DAP	146	68	2.15:1	0.230
Endosperm_15DAP	155	70	2.21:1	0.296
Endosperm_20DAP	157	69	2.28:1	0.372
Leaf1_2week	126	74	1.70:1	1.32×10^{-2}
Leaf2_2week	102	83	1.23:1	1.98×10^{-5}
Leaf1_3week	122	71	1.72:1	1.65×10^{-2}
Leaf2_3week	101	74	1.36:1	3.16×10^{-4}
Leaf3_3week	92	76	1.21:1	2.44×10^{-5}
Leaf_4week	88	85	1.04:1	2.81×10^{-7}
Leaf_5week	77	82	0.94:1	3.29×10^{-8}
Leaf_6week	112	67	1.67:1	1.27×10^{-2}
Leaf_9week	116	85	1.36:1	1.67×10^{-4}
Leaf_expanded_4week	135	89	1.52:1	1.13×10^{-3}
Leaf_immature_5week	120	84	1.43:1	4.41×10^{-4}
Leaf_mature_5week	138	85	1.62:1	4.43×10^{-3}
Leaf_wrapped_4week	116	87	1.33:1	9.11×10^{-5}
Pollen_early	89	66	1.35:1	4.24×10^{-4}
Pollen_germination	54	45	1.20:1	4.00×10^{-4}
Pollen1_mature	70	49	1.43:1	3.43×10^{-3}
Pollen2_mature	84	46	1.83:1	4.31×10^{-2}
Root1_2week	96	79	1.22:1	2.09×10^{-5}
Root2_2week	112	66	1.70:1	1.62×10^{-2}
Seed	123	68	1.81:1	3.56×10^{-2}
Shoot apical meristem_4week	81	67	1.21:1	5.08×10^{-5}
Stalk_2week	103	81	1.27:1	4.81×10^{-5}
Silk_9week	64	78	0.82:1	1.77×10^{-9}
Silk_11week	115	84	1.37:1	1.88×10^{-4}
Silk_12week	101	89	1.13:1	1.92×10^{-6}
Tassel_4week	101	76	1.33:1	1.66×10^{-4}
Tassel_8week	65	58	1.12:1	2.93×10^{-5}
Tassel_9week	100	91	1.10:1	7.21×10^{-7}
Vegetative apex_2week	72	64	1.12:1	1.64×10^{-5}

[a] Singleton-to-duplicate ratios.

[b] P values were calculated using the χ^2 goodness of fit test to compare the ratios of singletons to duplicates between miRNA genes in the whole genome and those expressed in different tissues.

specific *miR2275* family member was evolved as causative locus for the phenotypic transition in anthers by artificial selection.

Differentiation of 21- and 24-Nt *PHAS* Loci in Relation to Tissue Specificity

The 22-nt mature sequences of *miR2275* family members were recognized as critical triggers for 24-nt *PHAS* loci to generate phasiRNAs (Fei et al., 2013). In addition to the *miR2275* family, there is growing evidence supporting the tissue specificity of numerous *PHAS* triggers, such as *miR156*, *miR529*, and *miR2118*. Therefore, we further examined whether the expression patterns of *PHAS* loci are tissue-specific in maize. To address this question, a genome-wide identification of *PHAS* loci was performed using sRNA-seq data collected from 14 tissues. A total of 469 21-nt and 190 24-nt *PHAS* loci were identified with high confidence ($P \leq 10^{-5}$), respectively (**Supplementary Table 11**, **Supplementary Data 2**, and Supplementary Data 3). Genome-wide identification of *PHAS* loci in maize W23 inbred line was previously performed by Zhai et al (Zhai et al., 2015). In our study, 375 21-nt *PHAS* loci and 113 24-nt *PHAS* loci were overlapped with those identified by Zhai et al, respectively (**Supplementary Table 11**). Meanwhile, 94 21-nt *PHAS* loci and 77 24-nt *PHAS* loci were novel. Integrating the results from two studies together, we found that the two classes of *PHAS* loci were highly expressed in reproductive tissues (TPM ≥5; **Figure 6A**). The number of *PHAS* loci expressed in individual tissues varied from 0 to 580, indicating tissue-specific expression (**Supplementary Table 12**). To gain insight into the contribution of tissue type on the production of phasiRNAs, the accumulations of phasiRNAs and corresponding trigger miRNAs were compared in individual tissues. The accumulations of phasiRNAs were positively correlated with the expression levels of corresponding trigger miRNAs (**Figure 6A**). The 21- and 24-nt phasiRNAs exhibited strikingly different spatiotemporal regulation. Compared with 21-nt phasiRNAs, the 24-nt phasiRNAs were preferentially expressed in pollen tissues and suppressed in early anther development (**Figure 6A**).

Then we analyzed the genomic distribution of *PHAS* loci and found that the ratio of 21-nt *PHAS* loci in TEs to those in PCGs (2.94:1) was significantly higher than that of 24-nt *PHAS* loci (1.85:1; χ^2 test, $P < 0.05$). This observation indicated evolutionary divergence of 21- and 24-nt *PHAS* loci, which was also reflected by the distinct expression patterns of the two categories *PHAS* loci in **Figure 6B**. Although the singleton-to-duplicate ratio of 21-nt *PHAS* loci (2.34:1) was approximately consistent with that of 24-nt *PHAS* loci (2.21:1), there were obvious differences between the expression patterns of 21- and 24-nt *PHAS* loci (**Figure 6B**).

Taken together, these results suggested that divergence on local genomic environments and expression patterns occurred between 21- and 24-nt *PHAS* loci, appearing to promote the formation of intraspecific homogenization and interspecific heterogenization in terms of tissue-specific function.

DISCUSSION

Comprehensive and High-Quality Annotation of miRNA Genes in Maize

It is increasingly recognized that miRNAs are key components of gene regulation, and a diversity of mechanisms of action have been proposed (Yu et al., 2017). However, the evolutionary implications of maize miRNAs on tissue specificity are still uncharacterized. Studies of miRNA genes have been frequently confounded by questionable miRNA annotations and a lack of

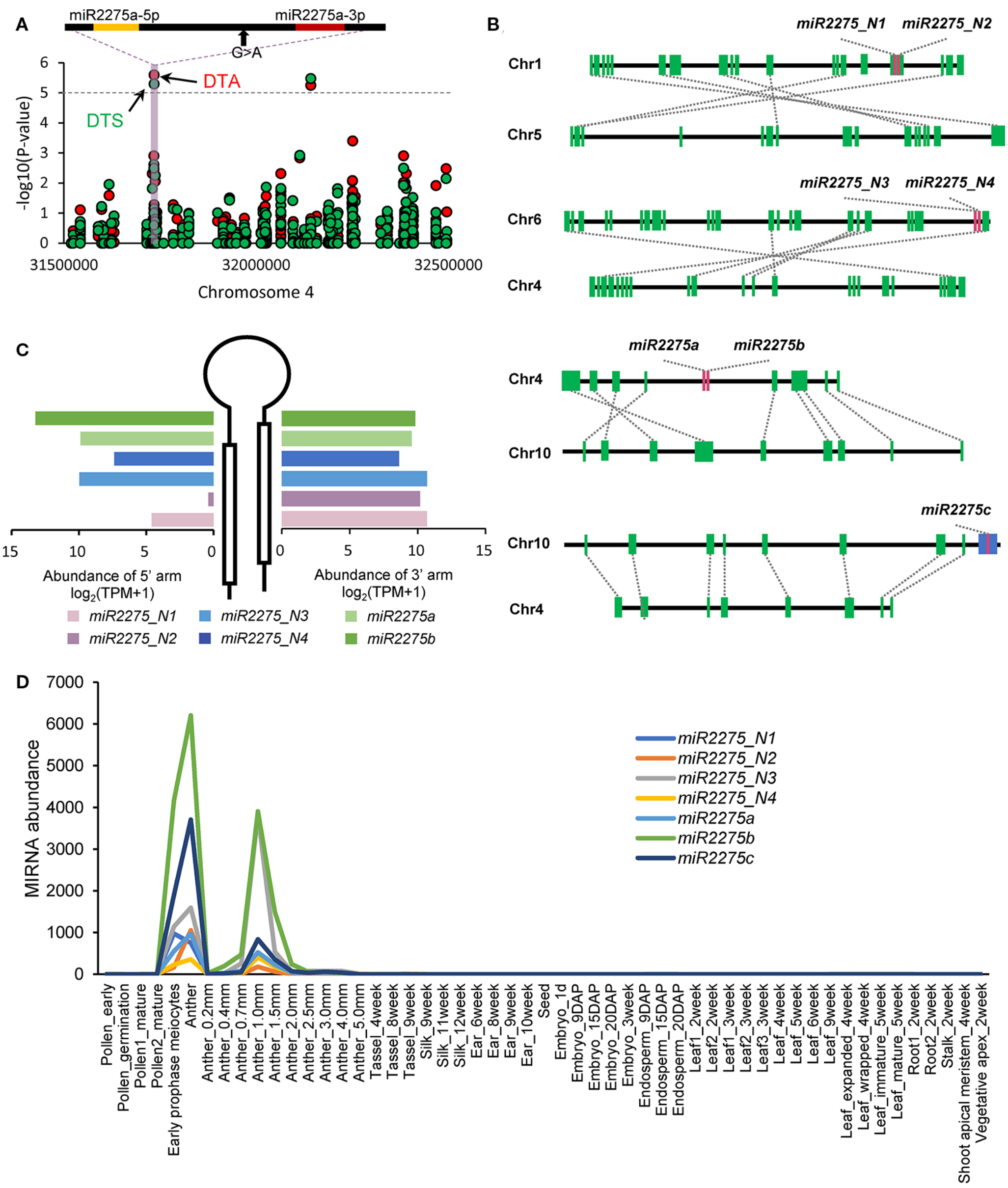


FIGURE 5 | Evolutionary consequence of the *miR2275* family in maize. **(A)** GWAS results, which overlapped with *miR2275* loci. The gray horizontal dashed lines indicate the significance threshold of GWAS (1×10^{-5}). The mature miRNA sequences are highlighted in red and the miRNA star sequences are highlighted in yellow. Black arrows indicate significant mutation sites. Days to anthesis (DTA) and days to silk production (DTS) are labeled in red and green, respectively. **(B)** Duplication status of *miR2275* family members. Green boxes indicate PCGs, red boxes represent *miR2275* gene family members, and the blue box represents helitron transposon. **(C)** Comparison of mature miRNA accumulation between 5′ and 3′ arms in individual *miR2275* members. **(D)** Accumulation levels of seven *miR2275* genes in individual tissue samples. The accumulation levels were calculated as sum of miRNA mature sequences, miRNA star sequences and isomiRs.

standardized criteria for the identification of plant miRNAs from the sRNA-seq datasets. In addition, miRNA genes are highly tissue specific, making it difficult to detect accurate abundance levels, unless massive amounts of tissues are available. In our study, we addressed these shortcomings by integrating 195 deeply sequenced sRNA libraries and by surveying 14 tissues in maize. This large-scale analysis was performed based on the recently updated criteria which can simultaneously emphasize accuracy and minimize false positives in the genome-wide annotation of plant miRNAs (Axtell and Meyers, 2018; Kuang et al., 2018). To our knowledge, this collection of sRNA-seq datasets is the largest effort to date for the genome-wide identification of miRNA genes in maize.

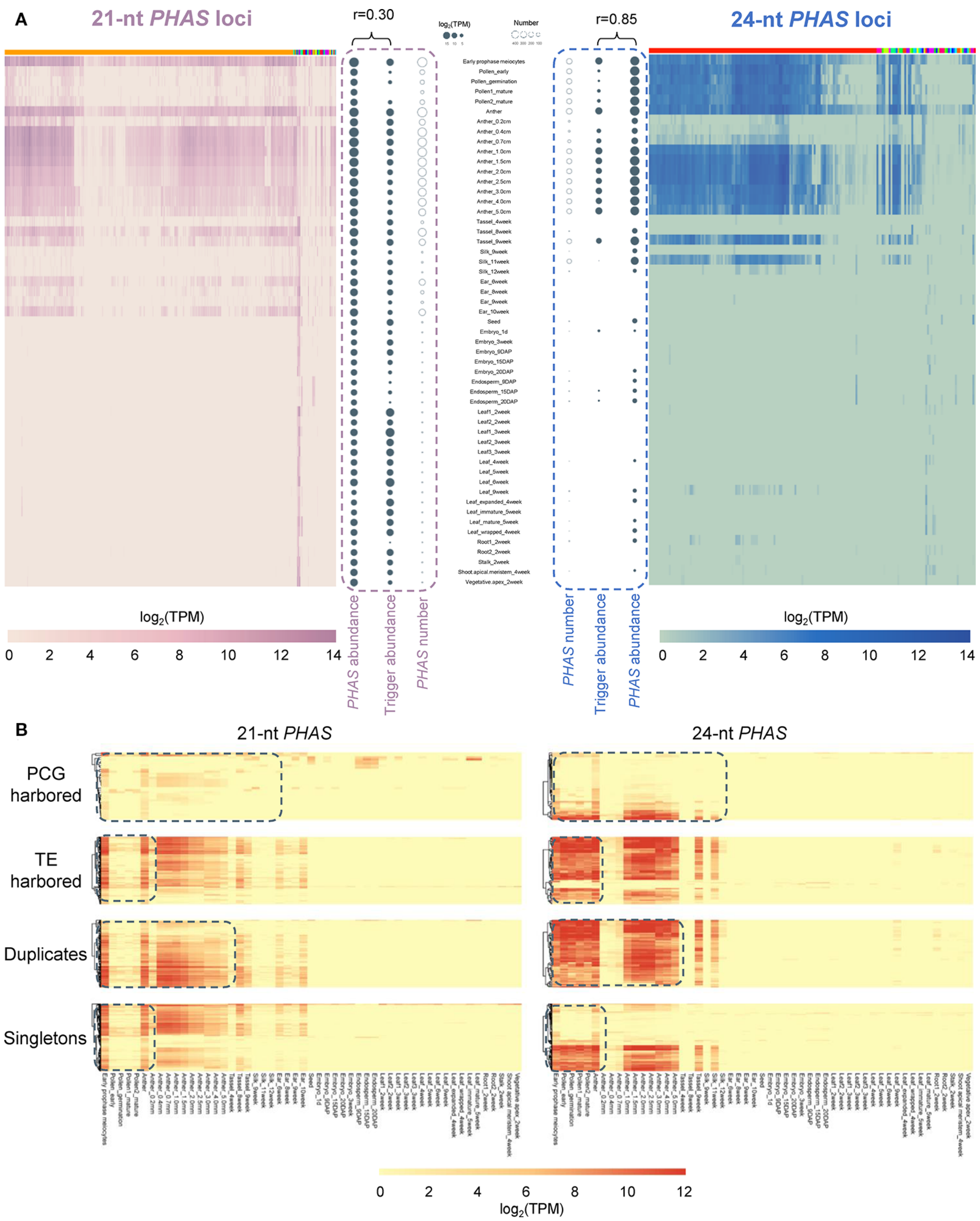


FIGURE 6 | Differentiation of 21- and 24-nt *PHAS* loci in relation to tissue specificity. **(A)** Heat maps depicted the abundance of 21-nt *PHAS* loci (left) and 24-nt *PHAS* loci (right) in individual tissue samples. Solid bubbles represent the total abundance of phasiRNAs and corresponding miRNA triggers. Circles represent the number of *PHAS* loci. In the heatmaps, *PHAS* loci were clustered according to the miRNA triggers. Color bars on the top of heatmaps represent different triggers, among which orange represents miR2118 and red represents miR2275. **(B)** Heat maps depicted the abundance of 21- and 24-nt *PHAS* loci in four categories (PCG harbored, TE harbored, duplicates, and singletons). The dashed boxes indicated the divergence between the abundance of 21- and 24-nt *PHAS* loci in corresponding tissue samples.

miRNAs as gene regulators are reported to participate in the regulation of drought responsive PCGs in many plant species (Ferdous et al., 2015). For instance, *miR162* (Tian et al., 2015) and *miR164* (Fang et al., 2014) are induced by drought stress in rice. Our study confirmed this regulation that, to our knowledge, had not been reported in maize. This implied functional conservation of these miRNAs between rice and maize. However, some of the drought-responsive miRNAs in plants, such as *miR319* (Zhou et al., 2013), *miR393* (Bian et al., 2012), and *miR396* (Liu et al., 2009), had very low abundance in our samples and relatively stable expressions under drought stress. This is probably because the promoters of these miRNA genes are not active in the tissues analyzed in our experiment. Nevertheless, we cannot rule out the possibility that these miRNAs do not function in response to drought stress in maize.

Local Genomic Stability Contributes to the Tissue Specificity of miRNAs and Variants

Local genomic features, such as the density of PCGs, proportion of TEs and rates of recombination, affect the removal of both genic and TE sequences (Wicker et al., 2007; Tian et al., 2009). We demonstrated that most of maize miRNA genes were distributed in chromosome arms, thus reflecting a preferential evolution driven by genetic recombination. The disruptive effects of TEs have been documented extensively, as they integrate into the regulatory or coding region of host genes or induce ectopic/nonallelic recombination (Jangam et al., 2017). We observed that there were significant differences among expression levels of miRNAs located in PCGs, TEs, and UIs. Although TEs constitute approximately 90% of the maize genome (SanMiguel et al., 1996), only 31.1% of miRNA genes were located in TEs. Such a low retention rate in TEs is more likely due to a faster purging rate of miRNA genes from TEs, which generally has limited deleterious effects on gene function and is, thus, more easily purified. We observed an extremely high abundance of miRNA genes in UIs compared with PCGs and TEs. If UIs are nearly neutral, the detected higher abundance for miRNA genes in UIs indicated that these intergenic sequences are under some level of specific functional constraints.

The modification of miRNAs, which gives rise to many isomiRs, occurs extensively and affects miRNA diversity or functional specificity (Sablok et al., 2015). Our analysis demonstrated that the number of isomiRs and frequency of isomiR types varied dramatically in different tissues. The first nucleotides on both 5′ and 3′ ends of miRNAs were generally ‘U’. It is known that 3′ adenylation increases miRNA stability, and 3′ uridylation enhances miRNA degradation in plants (Lu et al., 2009; Wang et al., 2019). Thus, the presence of ‘U’ on the terminus of 5′ addition isomiRs in all maize tissues suggested that the modification of isomiRs probably occurred to avoid degradation. Additionally, we observed diverse terminal nucleotides on the 3′ end of isomiRs among the individual maize tissues, implying a tissue-specific development landmark, where the activities of enzymes involved in miRNA degradation or modification are dramatically adjusted. The abundance of isomiRs was highly associated with the genomic location of miRNA genes, implying that local genomic stability might be a principal factor contributing to the evolution of tissue specificity among maize isomiRs.

Complex Interplays May Exist Among Duplication Status, Local Genome Stability, Tissue Specificity, and miRNA Abundance

Gene duplication is a major factor responsible for the evolution of genes with novel functions. When both members of a duplicated gene pair are retained, it is believed that neither member was involved in local genomic recombination. By contrast, a singleton originates solely from a local deletion, insertion, or translocation (Panchy et al., 2016). In general, PCGs often accumulate in genomic regions with active recombination. In this study, we found that the singleton-to-duplicate ratio of miRNA genes in PCGs was significantly higher than that in TEs and UIs, suggesting that miRNA singletons located in PCGs were likely formed through the accumulation of small deletions generated via genetic recombination, which removed single gene. These observations suggested that most miRNA singletons in maize, particularly those located in PCGs, were formed by the rapid removal of their duplicate copies after neofunctionalization. Like miRNA genes in PCGs, relatively high rates of preservation of singletons were also observed in tissue-specific miRNAs. This indicated that relatively stronger selection acted on tissue-specific miRNA singletons, which contributed to functional constraints associated with tissue specificity.

In addition, we observed that the singleton-to-duplicate ratios of expressed miRNAs in individual tissues were significantly lower than the whole genome-wide ratio. This implied a relatively higher number of non-expressed miRNA singletons than duplicates. Together with the higher genome-wide abundance miRNA duplicates than singletons, these phenomena support the gene balance hypothesis, according to which a successful genome has evolved an optimum balance of gene products binding with one another to produce multi-subunit complexes. If a gene pair is fractionated, dosage imbalance may reduce fitness or have lethal consequences. Thus, gene duplicates are generally retained (Freeling, 2009). We found that the singleton-to-duplicate ratio of miRNA genes in PCGs was significantly higher than those in TEs and UIs, and in PCGs, the accumulations of miRNA singletons were significantly lower than those of miRNA duplicates. These observations implied that the ratio of miRNA singletons in PCGs may be affected largely by gene dosage of the resident PCGs, and highly expressed miRNA genes were preferentially retained as duplicates.

GD Events Have Effects on Evolutionary Dynamics of Maize miRNA Loci in Relation to Tissue Specificity

Gene families develop via duplication of an original ancestral gene, and gene duplicates subsequently diverge from the original gene under selection (Chauve et al., 2008). This divergence occurs at varied rates, depending on the selective forces. In the case of *miR164* and *miR399* families, gene duplication and divergence appeared to result in the sub-functionalization of members of these two families. Although both families were reported to function in the development of maize roots (Tang and Chu, 2017; Hashimoto et al., 2018), we showed that *miR164a* and *miR164d* were specifically expressed in silk, and *miR399b* and *miR399e* exhibited leaf-specific expression. These two cases are typical examples of sub-functionalization, where gene copies originated from an ancestral gene eventually acquire specialized function in different tissues. Often, the functional diversification of miRNA gene duplicates occurs over time through a series of modifications to archetype miRNAs, such as nucleotide polymorphisms, resulting in the specialization of miRNA gene function in a particular tissue (Ameres and Zamore, 2013). This typically involves changes in complementary pairing with target genes that drive the expression dynamics in relevant pathways (Baldrich et al., 2018). In animals, the seed region is considered as the basic determinant of miRNA specificity (Chandradoss et al., 2015). Here, we showed that despite the divergence in the expression profiles of *miR164* and *miR399* family members, the seed region was considerably conserved (**Supplementary Figure 8**). Based on these observations, we proposed that SNPs preferentially occurred in the 3′ tail region of miRNA genes after GD, thus leading to the functional divergence of each duplicate copy. In the case of *miR2275* family, of which one member showed anther-specific expression, three homoeologous pairs formed by TD were detected. However, their syntenic copies were absent from the maize genome. Together with their consistent expression profiles, we hypothesized that paralogs produced by the WGD event lost function and were eventually deleted prior to the TD event.

Further Perspectives: Exploring the Dynamics of Newly Identified miRNAs in Different Maize Tissues With Spatial, Temporal, and Environmental Dimensions

In this study, we newly identified 271 maize miRNA genes and uncovered the tissues-specific expression patterns. As important regulatory elements of PCGs, miRNAs have gained increasing interests in the last few years to understand their dynamics roles of regulation mechanism involved in a variety of biological processes associated with growth, development, and stress responses in plants. We expect that, in the future, these new miRNA genes will be explored using large-scale sRNA-seq data from different developmental stages and tissues of maize under diverse environmental conditions. Moreover, further in-depth structural and functional analysis using genetic and molecular biology technologies will be performed to validate these new miRNA genes in maize.

DATA AVAILABILITY STATEMENT

The six sRNA-seq datasets generated in our study have been deposited into the NCBI's SRA database under the accession numbers SRR6322649, SRR6322650, SRR6322651, SRR6322652, SRR6322653, and SRP125654.

AUTHOR CONTRIBUTIONS

YX and TZ performed the research and analyzed the data. YL contributed analytic tools and methods. ZM conceived the project and wrote the article.

FUNDING

This work was supported by the Natural Science Foundation Research Project of Shaanxi Province of China (2019JQ-096), the College Students' Platform for Innovation Training Program from Northwest A&F University (201803094), and the Fund of Northwest A&F University (Z111021603).

ACKNOWLEDGMENTS

We thank Chuang Ma at Northwest A&F University for his discussion on the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00051/full#supplementary-material

REFERENCES

Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25 (1), 130–131. doi: 10.1093/bioinformatics/btn604

Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* 121 (2), 207–221. doi: 10.1016/j.cell.2005.04.004

Ameres, S. L., and Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* 14 (8), 475–488. doi: 10.1038/nrm3611

Audic, S., and Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res.* 7 (10), 986–995. doi: 10.1101/gr.7.10.986

Axtell, M. J., and Meyers, B. C. (2018). Revisiting criteria for plant MicroRNA annotation in the era of big data. *Plant Cell* 30 (2), 272–284. doi: 10.1105/tpc.17.00851

Baldrich, P., Beric, A., and Meyers, B. C. (2018). Despacito: the slow evolutionary changes in plant microRNAs. *Curr. Opin. Plant Biol.* 42, 16–22. doi: 10.1016/j.pbi.2018.01.007

Bian, H., Xie, Y., Guo, F., Han, N., Ma, S., Zeng, Z., et al. (2012). Distinctive expression patterns and roles of the miRNA393/TIR1 homolog module in regulating flag leaf inclination and primary and crown root growth in rice (Oryza sativa). *New Phytol.* 196 (1), 149–161. doi: 10.1111/j.1469-8137.2012.04248.x

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17, 66. doi: 10.1186/s13059-016-0924-1

Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., MacRae, I. J., and Joo, C. (2015). A dynamic search process underlies microRNA targeting. *Cell* 162 (1), 96–107. doi: 10.1016/j.cell.2015.06.032

Chauve, C., Doyon, J. P., and El-Mabrouk, N. (2008). Gene family evolution by duplication, speciation, and loss. *J. Comput. Biol.* 15 (8), 1043–1062. doi: 10.1089/cmb.2008.0054

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., et al. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.* 33 (20), e179. doi: 10.1093/nar/gni178

Creasey, K. M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B. C., et al. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature* 508 (7496), 411–415. doi: 10.1038/nature13069

Dai, X., and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39 (Web Server issue), W155–W159. doi: 10.1093/nar/gkr319

Doebley, J. (2004). The genetics of maize evolution. *Annu. Rev. Genet.* 38, 37–59. doi: 10.1146/annurev.genet.38.072902.092425

Edwards, J. W., and Coruzzi, G. M. (1990). Cell-specific gene expression in plants. *Annu. Rev. Genet.* 24, 275–303. doi: 10.1146/annurev.ge.24.120190.001423

Fahlgren, N., and Carrington, J. C. (2010). miRNA target prediction in plants. *Methods Mol. Biol.* 592, 51–57. doi: 10.1007/978-1-60327-005-2_4

Fang, Y., Xie, K., and Xiong, L. (2014). Conserved miR164-targeted NAC genes negatively regulate drought resistance in rice. *J. Exp. Bot.* 65 (8), 2119–2135. doi: 10.1093/jxb/eru072

Fei, Q., Xia, R., and Meyers, B. C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell* 25 (7), 2400–2415. doi: 10.1105/tpc.113.114652

Ferdous, J., Hussain, S. S., and Shi, B. J. (2015). Role of microRNAs in plant drought tolerance. *Plant Biotechnol. J.* 13 (3), 293–305. doi: 10.1111/pbi.12318

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453. doi: 10.1146/annurev.arplant.043008.092122

Hashimoto, K., Miyashima, S., Sato-Nara, K., Yamada, T., and Nakajima, K. (2018). Functionally diversified members of the MIR165/6 gene family regulate ovule morphogenesis in Arabidopsis thaliana. *Plant Cell Physiol.* 59 (5), 1017–1026. doi: 10.1093/pcp/pcy042

Holub, E. B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. *Nat. Rev. Genet.* 2 (7), 516. doi: 10.1038/35080508

Hung, P. S., Chang, K. W., Kao, S. Y., Chu, T. H., Liu, C. J., and Lin, S. C. (2012). Association between the rs2910164 polymorphism in pre-mir-146a and oral carcinoma progression. *Oncol.* 48 (5), 404–408. doi: 10.1016/j.oraloncology.2011.11.019

Jangam, D., Feschotte, C., and Betran, E. (2017). Transposable element domestication as an adaptation to evolutionary conflicts. *Trends In Genet.* 33 (11), 817–831. doi: 10.1016/j.tig.2017.07.011

Jeong, D. H., Park, S., Zhai, J., Gurazada, S. G., De Paoli, E., Meyers, B. C., et al. (2011). Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell* 23 (12), 4185–4207. doi: 10.1105/tpc.111.089045

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546 (7659), 524–527. doi: 10.1038/nature22971

Kadota, K., Ye, J., Nakai, Y., Terada, T., and Shimizu, K. (2006). ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinf.* 7, 294. doi: 10.1186/1471-2105-7-294

Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S., et al. (2017). PHASIS: a computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. *bioRxiv*. 158832. doi: 10.1101/158832

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42 (Database issue), D68–D73. doi: 10.1093/nar/gkt1181

Kuang, Z., Wang, Y., Li, L., and Yang, X. (2018). miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* 35 (14), 2521–2522. doi: 10.1093/bioinformatics/bty972

Li, S., Castillo-Gonzalez, C., Yu, B., and Zhang, X. (2017). The functions of plant small RNAs in development and in stress responses. *Plant J.* 90 (4), 654–670. doi: 10.1111/tpj.13444

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, D., Song, Y., Chen, Z., and Yu, D. (2009). Ectopic expression of miR396 suppresses GRF target gene expression and alters leaf growth in Arabidopsis. *Physiol. Plant.* 136 (2), 223–236. doi: 10.1111/j.1399-3054.2009.01229.x

Liu, Q., Wang, H., Zhu, L., Hu, H., and Sun, Y. (2013a). Genome-wide identification and analysis of miRNA-related single nucleotide polymorphisms (SNPs) in rice. *Rice* 6 (1), 10. doi: 10.1186/1939-8433-6-10

Liu, Y., Wang, Y., Zhu, Q. H., and Fan, L. (2013b). Identification of phasiRNAs in wild rice (Oryza rufipogon). *Plant Signaling Behav.* 8 (8), e25079. doi: 10.4161/psb.25079

Lu, S., Sun, Y. H., and Chiang, V. L. (2009). Adenylation of plant miRNAs. *Nucleic Acids Res.* 37 (6), 1878–1885. doi: 10.1093/nar/gkp031

Lyu, J. (2017). Maize domestication: an ancient genome speaks. *Nat. Plants* 3, 16215. doi: 10.1038/nplants.2016.215

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. *Genome Res.* 16 (4), 510–519. doi: 10.1101/gr.4680506

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., et al. (2008). Criteria for annotation of plant MicroRNAs. *Plant Cell* 20 (12), 3186–3190. doi: 10.1105/tpc.108.064311

Panchy, N., Lehti-Shiu, M., and Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171 (4), 2294–2316. doi: 10.1104/pp.16.00523

Petsch, K., Manzotti, P. S., Tam, O. H., Meeley, R., Hammell, M., Consonni, G., et al. (2015). Novel DICER-LIKE1 siRNAs bypass the requirement for DICER-LIKE4 in maize development. *Plant Cell* 27 (8), 2163–2177. doi: 10.1105/tpc.15.00194

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi: 10.1086/519795

Sablok, G., Srivastva, A. K., Suprasanna, P., Baev, V., and Ralph, P. J. (2015). isomiRs: Increasing evidences of isomiRs complexity in plant stress functional biology. *Front. In Plant Sci.* 6, 949. doi: 10.3389/fpls.2015.00949

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274 (5288), 765–768. doi: 10.1126/science.274.5288.765

Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009). CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.* 37 (Web Server issue), W277–W280. doi: 10.1093/nar/gkp367

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protools* 3 (6), 1101–1108. doi: 10.1038/nprot.2008.73

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* 108 (10), 4069–4074. doi: 10.1073/pnas.1101368108

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., et al. (2009). SNPs in human miRNA genes affect biogenesis and function. *RNA* 15 (9), 1640–1651. doi: 10.1261/rna.1560209

Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., et al. (2004). Close split of sorghum and maize genome progenitors. *Genome Res.* 14 (10A), 1916–1923. doi: 10.1101/gr.2332504

Tang, J., and Chu, C. (2017). MicroRNAs in crop improvement: fine-tuners for complex traits. *Nat. Plants* 3, 17077. doi: 10.1038/nplants.2017.77

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., et al. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* 19 (12), 2221–2230. doi: 10.1101/gr.083899.108

Tian, C., Zuo, Z., and Qiu, J. L. (2015). Identification and characterization of ABA-responsive microRNAs in rice. *J. Genet. Genomics* 42 (7), 393–402. doi: 10.1016/j.jgg.2015.04.008

Tu, B., Liu, L., Xu, C., Zhai, J., Li, S., Lopez, M. A., et al. (2015). Distinct and cooperative activities of HESO1 and URT1 nucleotidyl transferases in microRNA turnover in Arabidopsis. *PloS Genet.* 11 (4), e1005119. doi: 10.1371/journal.pgen.1005119

Voorrips, R. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93 (1), 77–78. doi: 10.1093/jhered/93.1.77

Wang, J., Mei, J., and Ren, G. (2019). Plant microRNAs: biogenesis, homeostasis, and degradation. *Front. Plant Sci.* 10, 360. doi: 10.3389/fpls.2019.00360

Wicker, T., Yahiaoui, N., and Keller, B. (2007). Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. *Plant J.* 51 (4), 631–641. doi: 10.1111/j.1365-313X.2007.03164.x

Wu, H. J., Ma, Y. K., Chen, T., Wang, M., and Wang, X. J. (2012). PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Res.* 40 (Web Server issue), W22–W28. doi: 10.1093/nar/gks554

Yu, Y., Jia, T., and Chen, X. (2017). The 'how' and 'where' of plant microRNAs. *New Phytol.* 216 (4), 1002–1017. doi: 10.1111/nph.14834

Zhai, J., Zhao, Y., Simon, S. A., Huang, S., Petsch, K., Arikit, S., et al. (2013). Plant microRNAs display differential 3' truncation and tailing modifications that are ARGONAUTE1 dependent and conserved across species. *Plant Cell* 25 (7), 2417–2428. doi: 10.1105/tpc.113.114603

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G. L., Walbot, V., et al. (2015). Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc. Natl. Acad. Sci. U. S. A.* 112 (10), 3146–3151. doi: 10.1073/pnas.1418918112

Zhao, M., Meyers, B. C., Cai, C., Xu, W., and Ma, J. (2015). Evolutionary patterns and coevolutionary consequences of MIRNA genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean. *Plant Cell* 27 (3), 546–562. doi: 10.1105/tpc.15.00048

Zhong, X., Pla, A., and Rayner, S. (2019). Jasmine: a Java pipeline for isomiR characterization in miRNA-seq Data. *Bioinformatics* btz806 doi: 10.1093/bioinformatics/btz806

Zhou, M., Li, D., Li, Z., Hu, Q., Yang, C., Zhu, L., et al. (2013). Constitutive expression of a miR319 gene alters plant development and enhances salt and drought tolerance in transgenic creeping bentgrass. *Plant Physiol.* 161 (3), 1375–1391. doi: 10.1104/pp.112.208702

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 March 2020
doi: 10.3389/fgene.2020.00178

Construction of a SNP-Based Genetic Map Using SLAF-Seq and QTL Analysis of Morphological Traits in Eggplant

*Qingzhen Wei, Wuhong Wang, Tianhua Hu, Haijiao Hu, Jinglei Wang and Chonglai Bao**

Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou, China

Eggplant (*Solanum melongena*; 2*n* = 24) is an economically important fruit crop of the family Solanaceae that was domesticated in India and Southeast Asia. Construction of a high-resolution genetic map and map-based gene mining in eggplant have lagged behind other crops within the family such as tomato and potato. In this study, we conducted high-throughput single nucleotide polymorphism (SNP) discovery in the eggplant genome using specific length amplified fragment (SLAF) sequencing and constructed a high-density genetic map for the quantitative trait locus (QTL) analysis of multiple traits. An interspecific F_2 population of 121 individuals was developed from the cross between cultivated eggplant "1836" and the wild relative *S. linnaeanum* "1809." Genomic DNA extracted from parental lines and the F_2 population was subjected to high-throughput SLAF sequencing. A total of 111.74 Gb of data and 487.53 million pair-end reads were generated. A high-resolution genetic map containing 2,122 SNP markers and 12 linkage groups was developed for eggplant, which spanned 1530.75 cM, with an average distance of 0.72 cM between adjacent markers. A total of 19 QTLs were detected for stem height and fruit and leaf morphology traits of eggplant, explaining 4.08–55.23% of the phenotypic variance. These QTLs were distributed on nine linkage groups (LGs), but not on LG2, 4, and 9. The number of SNPs ranged from 2 to 11 within each QTL, and the genetic interval varied from 0.15 to 10.53 cM. Overall, the results establish a foundation for the fine mapping of complex QTLs, candidate gene identification, and marker-assisted selection of favorable alleles in eggplant breeding.

Keywords: SLAF sequencing, SNP markers, eggplant, genetic linkage map, QTL analysis

OPEN ACCESS

Edited by:
Chuang Ma,
Northwest A&F University, China

Reviewed by:
Changwei Shao,
Yellow Sea Fisheries Research Institute (CAFS), China
Heena Ambreen,
National Institute of Plant Genome Research (NIPGR), India

****Correspondence:***
Chonglai Bao
baocl@mail.zaas.ac.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *22 October 2019*
Accepted: *13 February 2020*
Published: *11 March 2020*

Citation:
Wei Q, Wang W, Hu T, Hu H, Wang J and Bao C (2020) Construction of a SNP-Based Genetic Map Using SLAF-Seq and QTL Analysis of Morphological Traits in Eggplant. Front. Genet. 11:178. doi: 10.3389/fgene.2020.00178

INTRODUCTION

Eggplant (*Solanum melongena* L., 2*n* = 24) is an important vegetable crop cultivated worldwide that belongs to the large family Solanaceae. The total global production of eggplants accounted for ∼52.3 million tons in 2017 (FAOSTAT 2017[1]). In contrast to many New World-originating vegetables within the family, such as tomato, potato, and pepper, eggplant most likely originated in the Old World (Daunay et al., 2001). Eggplants exhibit extensive variations in leaf morphology, fruit size and shape, and plant architecture among cultivated varieties and wild relatives.

[1] http://faostat3.fao.org

However, map-based gene cloning, as well as the understanding of the molecular and genetic mechanisms underlying horticulturally important traits in eggplant, have largely lagged behind compared to other vegetable crops (i.e., tomato and cucumber) due to the limited number of molecular markers and relatively low density of genetic maps.

In the past decades, a number of eggplant linkage maps have been constructed with both dominant and co-dominant DNA markers, which have been used to map disease resistance and plant morphology traits (Nunome et al., 2001, 2003, 2009; Wu et al., 2009; Barchi et al., 2010, 2012, 2018; Lebeau et al., 2013; Frary et al., 2014; Miyatake et al., 2016). The first eggplant genetic map was constructed by Nunome et al. (2001) using dominant markers [randomly amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) markers] and an intra-specific EWF_2 population; several markers were associated with fruit shape and color. However, the map contained only 181 markers and covered 21 linkage groups. Subsequently, Nunome et al. (2003, 2009) developed ~1000 simple sequence repeat (SSR) markers with which two genetic maps were constructed using the same F_2 population. Although these linkage maps are saturated, they have been useful as important co-dominant marker resources for marker-assisted selection in eggplant breeding. Miyatake et al. (2012) constructed two linkage maps containing SSRs and single nucletide polymorphisms (SNPs) and identified linked markers for parthenocarpy traits using NAF2 and ALF2 populations. To date, the most saturated genetic map was constructed by integrating the SSRs and SNPs with another two F_2 populations (LWF2 and EWF2); in total, 1,745 loci were mapped in the integrated map (Fukuoka et al., 2012; Hirakawa et al., 2014). This map was used to facilitate the assembly of eggplant draft genome sequences.

In parallel with the advances in the genetic linkage maps, the identification of quantitative trait locus (QTLs) associated with agronomic traits has been greatly promoted in eggplant. Doganlar et al. (2002a) constructed a molecular linkage map with tomato restriction fragment length polymorphism (RFLPs) and an interspecific F_2 population (58 individuals). A total of 125 significant QTLs associated with domestication and morphological traits were detected using the interspecific linkage map in the same segregation population (Doganlar et al., 2002b; Frary et al., 2003; Wu et al., 2009). Sunseri et al. (2003) developed an RAPD-AFLP map with 273 markers and identified molecular markers linked with *Verticillium* wilt. Barchi et al. (2010, 2012) identified a number of QTLs controlling anthocyanin pigmentation using an intra-specific F_2 population derived from 305E40 × 67/3 and a 238-loci linkage map. With the same linkage map and F_2 population, QTLs related to total and early yield, fruit traits (e.g., weight, length, diameter, and shape), prickliness traits, and fruit metabolic content were mapped, and, for each trait, at least one major QTL was identified (Portis et al., 2014; Toppino et al., 2016). Major and minor QTLs affecting resistance to *Fusarium* and *Verticillium* in the intraspecific 305E40 × 67/3 map were also detected, and putative orthologous genes from tomato were identified (Barchi et al., 2018). In addition, genome-wide association analysis (GWAS) also plays an important role in genetic mapping of relavent traits in eggplant. Using this approach, Cericola et al. (2014) and Portis et al. (2015) identified a number of phenotype/genotype associations for key breeding fruit and plant traits with 191 mixed eggplant accessions. Nonetheless, despite the progress in QTL detection, most of the traits were analyzed in intraspecific populations, and the linkage maps used in genetic mapping are still less saturated.

Large-scale DNA marker development and the construction of a high-resolution linkage map in eggplant would provide fundamental tools for map-based gene mining. Advances in next-generation sequencing (NGS) technologies provide an excellent opportunity to develop abundant SNP markers for linkage map construction. Restriction-site associated DNA sequencing (RAD-seq) and 2b-RAD are both useful tools for SNP discovery; they reduce genome complexity by sequencing only DNA fragments with restriction sites despite fragment length (Miller et al., 2007; Baird et al., 2008; Ogden, 2011; Wang et al., 2012). Specific length amplified fragment (SLAF) sequencing (SLAF-seq) is an improved reduced representation library (RRL) sequencing strategy that brings down the cost through genome reduction. SLAF-seq has been proved an effective method for *de novo* SNP discovery and high-throughput genotyping and has wide applications in genetic map construction in sesame (Zhang et al., 2013), kiwifruit (Huang et al., 2013), soybean (Qi et al., 2014; Zhang et al., 2018), cucumber (Wei et al., 2014; Zhu et al., 2016), peanut (Hu et al., 2018), sweet osmanthus (He et al., 2017), and cotton (Zhang et al., 2016; Ali et al., 2018). In the present study, we developed an interspecific F_2 population containing 121 individuals from a cross between an eggplant cultivar and the wild relative *S. linnaeanum*. Using the F_2 population and SLAF-seq technology, a high-density genetic map with 2,122 SNP markers was constructed. Importantly, 19 QTLs were detected for plant architecture-, fruit-, and leaf-related traits.

MATERIALS AND METHODS

Plant Materials and Phenotyping

The cultivated eggplant *S. melongena* "1836" and its wild relative *S. linnaeanum* "1809" were used as male and female parents, respectively. "1836" is an inbred line with long, purple fruits and few prickles, whereas "1809" is prickly and produces small, round, green, and striped fruit (**Figure 1**). In this study, an interspecific F_2 population containing 121 individuals was generated from a cross between "1809" and "1836," which was then used as the mapping population. The parents and the F_2 population were grown in spring 2018 in the greenhouses at Qiaosi experiment field of Zhejinag Academy of Agricultural Sciences, Hangzhou, China, with plant spacing of 60 cm, row spacing of 1.2 m, and ridge cultivation.

We collected data on the following eight traits from the F_2 plants: main stem height (MSH), fruit length (FL), fruit diameter (maximum diameter; FD), fruit shape (FL/FD; FS), leaf lobing (LLOB), leaf prickle number (LPN), leaf prickle color (LPC), and vein color (VC). The height of the main stem (in cm) is the length from above the ground to the pseudobinary branch, which was determined in Qiaosi in mid-July. Fruit character measurements

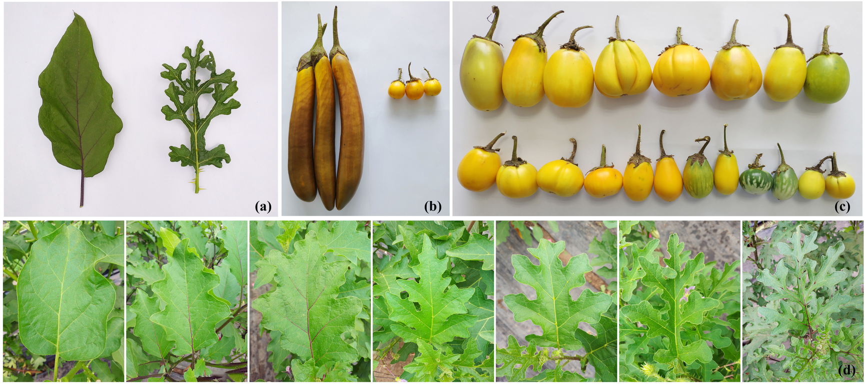


FIGURE 1 | Fruit and leaf morphology of the two eggplant parental lines and the F_2 population. **(a)** Leaves of *Solanum linnaeanum* "1809" (right) and *S. melongena* "1836" (left); **(b)** mature fruits of "1809" (right) and "1836" (left); **(c)** mature fruits of the representative F_2 individuals; **(d)** leaves of the representative F_2 individuals.

(in mm) were taken from mature fruits at the beginning of July. For fruit-related traits, three fruits were measured for each F_2 plant, and the average value was used as the final fruit trait value for QTL mapping. Plants with fewer than three fruits were considered as N/A, since some F_2 individuals could not produce fruits by self-pollination. Leaf lobing was assessed on a 1–7 scale (1 for no fissures at leaf margin; 7 for leaf margin deeply cleft); leaf prickle number was assessed in the same way (1 for no prickles; 7 for many prickles). Leaf prickle color was assessed on a 1-4 scale (light green, green-purple, green-brown, and dark purple). Leaf vein color was scored on a 1 to 3 scale (green, green and purple, and purple). All leaf morphology traits were measured at the beginning of July in Qiaosi.

DNA Extraction

Young, healthy leaves from both eggplant parents and the 121 F_2 individuals were collected, frozen in liquid nitrogen, and then transferred to a −80°C freezer. Total genomic DNA was extracted from each leaf sample according to the cetyltrimethyl ammonium bromide (CTAB) method (Murray and Thompson, 1980). DNA concentration and quality were examined by electrophoresis on 1% agarose gels using a standard lambda DNA and an ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, United States).

SLAF Library Construction and Sequencing

SLAF-seq was used to genotype the two parents and 121 F_2 individuals according to previously described procedures (Sun et al., 2013; Wei et al., 2014) with some modifications. The predicted electronic enzymatic digestion was performed using the eggplant genome sequence (Hirakawa et al., 2014) as a reference genome, and a combination of two endonucleases (*Rsa*I and *Hae*III) was selected. The digested fragment sequences with lengths of 414–464 bp were defined as a SLAF Label. The PCR products were purified using a Gel Extraction Kit (Qiagen), and then the gel-purified products were sequenced on an Illumina HiSeq 2500 system (Illumina, Inc., San Diego, CA, United States) according to the manufacturer's recommendations. The ratio of raw high-quality reads with quality scores greater than Q30 (a quality score of 30 indicates a 0.1% chance of obtaining an error, and thus 99.9% confidence) and the guanine-cytosine (GC) content were calculated for quality control.

SNP Discovery and Genotyping

SLAF-seq data were assessed using the software developed by Sun et al. (2013), and SNP markers were identified and genotyped according to the procedures described by Sun et al. (2013) and Wei et al. (2014) with modifications. Sequences with over 95% similarity were considered as one SLAF locus. The clean reads were mapped to the eggplant reference genome (Barchi et al., 2019), and the GATK software kit was used to detect SNPs between two parents and F_2 offspring. For the detailed process, see GATK's official website, https://www.broadinstitute.org/gatk/guide/best-practices.php (McKenna et al., 2010). SLAFs with two to four alleles were identified as polymorphic and considered potential markers. All polymorphism SLAFs were genotyped with consistency in the parental and offspring SNP loci. The polymorphic SLAF markers were sorted into eight segregation patterns as follows: ab × cd, ef × eg, hk × hk, lm × ll, nn × np, aa × bb, ab × cc, and cc × ab. Since the mapping populations were derived from two homozygous eggplant parents with a genotype of aa or bb, only the SLAF markers with the segregation pattern aa × bb were used.

Analysis of Segregation Distortion and Marker Filtering

To guarantee the quality of markers on the linkage map, we performed a strict marker filtering process according to the sequencing depth of the parents, marker completeness (coverage), and *P*-value of segregation distortion. Only SNPs with parental sequence depths of more than 10 × were retained, the complete degree parameters were set as 70 and 75%, and the *P*-values were set at less than 0.05, 0.01, and 0.001. The parameters were arranged in seven combinations to filter and analyze the markers.

According to the analysis of segregation distortion, we used two rules to filter the polymorphic markers in order to ensure the quality of the genetic map and the uniform distribution of markers. For linkage groups (LGs) 1, 3, 4, 6, 7, 8, 9, and 11, the following filtering principles were employed. (1) Filtering by sequence depth: remove SNPs with parental sequence depths of less than 10 ×. (2) Filtering by sequence completeness: remove SNPs with complete degree below 70%. (3) Filtering by segregation distortion: remove SNPs with serious segregation distortion (P-value < 0.01). Two additional filtering principles used for LGs 2, 5, 10, and 12 were as follows. (1) Filtering by sequence depth: remove SNPs with parental sequence depths of less than 12 ×. (2) Filtering by sequence completeness: remove SNPs with complete degree below 90%. According to these parameters, only high-quality SNPs were selected as potential markers.

Linkage Map Construction

The high-quality SNP markers were arranged and genotyping errors were corrected by the HighMap strategy and the SMOOTH algorithm, respectively (van Os et al., 2005; Liu et al., 2014). The k-nearest neighbor algorithm was used to manage missing genotypes (Huang et al., 2011). Then, the genetic linkage map was constructed by Joinmap v5.0[2], according to the regression algorithm method. The Kosambi mapping function was used to estimate the genetic distance (cM) of adjacent markers.

QTL Analysis

Quantitative trait locus analysis was performed with R/qtl[3]. Composite interval mapping (CIM) was used to identify QTLs. The logarithm of odds (LOD) threshold was used to evaluate the statistical significance of each QTL and was set by 1000 permutations test (PT). To ensure that both major and minor effect QTLs could be identified, different LOD scores were adopted. Firstly, a LOD threshold corresponding to 0.99 confidence was considered, and if there was no mapping interval, a LOD threshold corresponding to 0.95 confidence was used; if there was still no positioning interval, the threshold value of 0.90 confidence was considered. Finally, if there was still no QTL interval detected, the PT result was not used, and the threshold was manually lowered to 3.0, 2.5, and 2.0. QTLs were named according to their linkage group locations and trait names. For example, *fl1.1* referred to the first QTL for fruit length on eggplant LG 1.

RESULTS

SLAF-Sequencing and SNP Marker Analysis

After SLAF library construction and high-throughput sequencing, a total of 111.74 GB of data comprising 487.53 M paired-end reads was generated. Among these reads, 94.04% achieved or exceeded a quality score of 30 (Q30, indicating a 0.1% chance of an error, and 99.9% confidence) and the guanine-cytosine (GC) content was 38.83%. The average sequencing depth was 19.48 × for the female parent "1809," 19.46 × for the male parent "1836," and 10.29 × for F_2 progeny. GATK software was used to develop parent and offspring SNP markers combined with the eggplant reference genome. In total, 13,455,526 SNP markers were developed, in which 11,386,169 SNPs were successfully encoded and grouped into eight segregation patterns (ab × cd, ef × eg, hk × hk, lm × ll, nn × np, aa × bb, ab × cc, and cc × ab; **Supplementary Table S1**). 9,971,182 SNPs fell into the segregation pattern of aa × bb, accounting for 74.10% of the total developed markers, which were used for further analysis (**Supplementary Table S1**).

Segregation Distortion

According to the three parameters, i.e., parent sequencing depth, coverage degree, and P-value of segregation distortion, we calculated clustering results under seven parameters (**Table 1**), and the remaining number of markers was between 10,773 and 64,320. When sequencing depth and coverage were the same, marker number tended to increase as the P-value decreased. We found that LG2, 5, 10, and 12 had significantly lower marker numbers after segregation distortion filtering, regardless of which P-value was used. LG2 had the lowest number of SNPs, which ranged between 26 and 46 under different parameter combinations. The marker number on LG5 varied from 80 to 588; when sequencing depth was set > 10 × and coverage degree was 75%, 80 SNPs were retained with P-values < 0.05, whereas 577 SNPs were retained with P-values < 0.001; thus most markers on LG5 were segregation distorted. Similar results were also observed for LG10 and LG12 (**Table 1**). The marker numbers on the four linkage groups were slightly improved after decreasing sequencing depth, coverage degree, and P-value. However, the four linkage groups had a significantly higher number of markers if not selected for segregation distortion. Based on these results, we adopted two principles in the marker filter process to ensure both marker number and quality (See section "Materials and Methods").

Genetic Linkage Map

After further marker filtering and quality screening, a total of 2,122 SNP markers that met the quality standards were used for genetic map construction. The average depths of the SNPs for female, male, and the offspring were 15.03-fold, 14.09-fold, and 24.68-fold, respectively. JoinMap 5.0 assigned all of the 2,122 markers to 12 eggplant LGs (**Figure 2**), details of this SNP-based genetic map are presented in **Supplementary Table S1** and summarized in **Table 2**. The total genetic length of the eggplant linkage map was 1530.75 cM with an average marker distance of 0.72 cM. The number of SNP markers in each LG ranged from 90 (LG2) to 273 (LG9), with genetic distances spanning 38.67 cM (LG2) to 165.29 cM (LG1), and mean marker intervals ranged from 0.43 to 1.4 cM. The longest linkage group is LG1, which contains 167 SNPs, whereas the shortest is LG2, containing 90 SNPs.

[2] https://www.kyazma.nl/index.php/JoinMap/

[3] http://www.rqtl.org/

TABLE 1 | Number of SNP markers retained using different filtering parameters in 12 eggplant linkage groups.

Linkage group ID	Sequence Depth_Coverage_*P* value							
	10_0.7_0.05	10_0.7_0.01	10_0.7_0.001	10_0.75_0.05	10_0.75_0.01	10_0.75_0.001	10_0.75	12_0.1
E01	93	466	2339	93	465	2334	7010	2512
E02	26	32	46	26	32	46	5751	2186
E03	320	913	2300	320	913	2296	5103	1944
E04	1948	2286	2754	1948	2283	2744	5389	1959
E05	81	156	588	80	152	577	2268	757
E06	1993	3045	3585	1988	3028	3540	6336	2227
E07	2839	3487	4198	2830	3474	4173	8801	3171
E08	2435	3015	3554	2435	3003	3529	6787	2414
E09	497	592	686	497	589	679	1430	472
E10	34	68	177	33	67	175	6265	2484
E11	467	1132	1439	466	1130	1435	3303	1112
E12	40	66	106	40	66	105	5877	2397
Total	10773	15258	21772	10756	15202	21633	64320	23635

Phenotypic Evaluation

In the present study, the cultivated eggplant “1836” and the wild *S. linnaeanum* “1809” (**Figure 1**) were used to develop F_2 segregating populations for QTL analysis of multiple traits. Phenotypic data (including family means, standard errors, and distribution) of eight traits, i.e., main stem height (MSH), fruit length (FL), fruit diameter (FD), fruit shape (FS), leaf lobing (LLOB), leaf prickle number (LPN), leaf prickle color (LPC), and vein color (VC), are presented in **Supplementary Table S2**. All traits were measured in summer 2018. MSH, FL, FD, and FS could easily be measured or calculated using a ruler or vernier caliper; LLOB and LPN were scored on a 1-7 scale to describe the degree of lobing or number, respectively. Skewness and kurtosis tests showed that all these traits were normally distributed (**Supplementary Table S2**). LPC ranged in a spectrum from light green to dark purple in the segregating populations and was thus assessed on a 1-4 color scale. Likewise, VC was assessed on a 1-3 scale. They were also treated as quantitative traits. We calculated correlations among all eight traits (**Supplementary Table S2**). The results showed that LPC and VC had a notably high correlation and that FL was correlated closely with FD and FS. Moreover, considerable correlation was also observed between LLOB and LPN, with a correlation coefficient of 0.65. The highly correlated traits may share some tightly linked markers and/or candidate genes, such as the prickles on leaf and stem, fruit length and diameter, and the color of leaf vein and prickle. In fact, the genes involved in anthocyanin accumulation may simultaneously affect the colors of leaf vein, stem, prickle, and even fruit epicarp. This is useful in candidate gene prediction and function analysis. Fruit shape is determined by length and diameter; thus, breeding for eggplant fruit shape should take both into consideration.

QTL Mapping of Morphological Traits in Eggplant

A total of 19 QTLs for main stem height (*msh*), fruit length (*fl*), fruit diameter (*fd*), fruit shape (*fs*), leaf lobing (*llob*), leaf prickle number (*lpn*), leaf prickle color (*lpc*), and vein color (*vc*) were identified in the F_2 eggplant population (**Table 3**). The phenotypic variance explained (PVE) by the 19 QTLs ranged from 4.08-55.23%, and LOD values ranged from 2.09-28.75. The number of SNP markers within each QTL varied from 2 to 11, and the genetic distance interval of the QTLs ranged from 0.15 to 10.53 cM. The physical locations of these QTLs on eggplant chromosomes were obtained by BLAST marker sequences with the eggplant reference genome. The QTLs were distributed on nine chromosomes/linkage groups: LG1, 3, 5, 6, 7, 8, 10, 11, and 12 (**Figure 2**). Detailed sequence information and alignment positions for all markers are presented in **Table 3** and **Supplementary Table S3**. The SNPs of each QTL were mapped to the eggplant genome (Barchi et al., 2019) to anchor the physical locations, and the distribution of these QTL loci on chromosomes determined with two terminal markers are shown in **Supplementary Figure S1**.

One QTL for MSH, designated *msh5.1*, was detected on LG5 and could explain 6.8% of the observed phenotypic variation. Eleven SNPs were uncovered within this QTL region. However, although QTL *msh5.1* on the genetic map is a continuous interval, it corresponded to two physical regions on chromosome 5: 5.98-7.93 Mb and 13.57-19.22 Mb.

Four QTL loci, designated *fl1.1*, *fl5.1*, *fl7.1*, and *fl12.1*, were detected for FL and were located on LG1, 5, 7, and 12, respectively. The most prominent QTL, *fl12.1*, explained 24.05% of the phenotypic variation, followed by *fl1.1* and *fl5.1*, which explained 12.00 and 13.06% of the fruit length variation, respectively. The lowest contribution rate was 4.08% for QTL *fl7.1*. In addition, up to 11 SNPs were identified within the QTL region of *fl5.1*. Among the four QTLs, except that Marker8964317 of *fl12.1* is far from the other two markers, all other markers are within a reasonable genome position (**Table 3** and **Supplementary Table S3**). For *fl12.1*, the three markers within the region were positioned apart on chromosome 12, at 18.05, 31.63, and 90.73 Mb; the interval 18.05-31.63 Mb were retained since the other site, 90.73 Mb, was far from the other two sites.

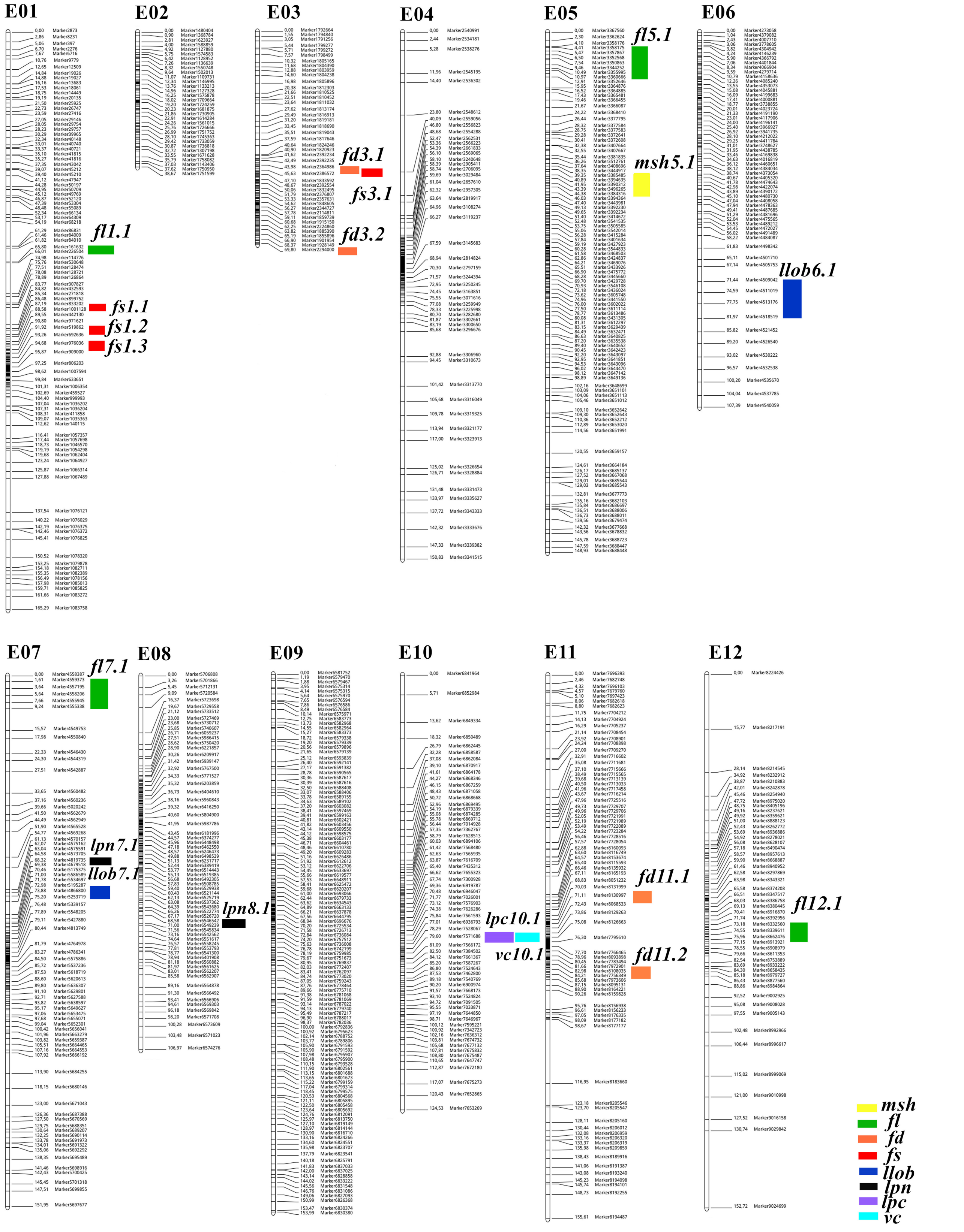


FIGURE 2 | SNP-based genetic linkage map of the interspecific F_2 population showing positions of QTLs. The numbers to the right of each LG indicate genetic distance (cM) between adjacent markers. The color bars refer to QTLs detected for the eight traits.

TABLE 2 | Information on the SNP-based genetic map of eggplant.

Linkage group ID	Marker no.	Total distance (cM)	Average distance (cM)
E01	167	165.29	0.99
E02	90	38.67	0.43
E03	164	73.89	0.45
E04	206	150.83	0.73
E05	219	148.93	0.68
E06	138	107.39	0.78
E07	166	151.95	0.92
E08	224	106.97	0.48
E09	273	153.97	0.56
E10	147	124.53	0.85
E11	219	155.61	0.71
E12	109	152.72	1.40
Total	2122	1530.75	0.72

For fruit diameter, four QTLs were detected on linkage groups 3 and 11, two for each LG. The highest contribution rate was 11.98% for *fd3.1*, followed by *fd3.2* (8.50%), *fd11.1* (7.95%), and *fd11.2* (6.62%). The genetic interval of each QTL ranged from 0.15 to 0.68 cM. The case in *fl12.1* was also observed for *fd3.2. fd11.1* and *fd11.2* spanned 71.47-71.87 cM and 82.29-82.98 cM on linkage group 11, respectively. Among the four SNP markers corresponding to *fd11.2*, Marker8108035 and Marker8108041 were at the same position of 55.83 Mb whereas the other two markers covered an interval of 13.15-28.19 Mb.

Four QTLs were detected for the FS index, with the contribution rates ranging from 6.92% (*fs1.1*) to 10.92% (*fs3.1*). There were three QTLs on LG1 (*fs1.1*, *fs1.2*, *fs1.3*), and the genetic locations were relatively close together. The chromosomal locations compared with *fs1.3* (111.30-111.91 Mb) were within the chromosomal location region of *fs1.3* (110.94-124.29 Mb). The other QTL was located on LG3, and its position was close to that of *fd3.1*.

The LLOB was controlled by two QTLs, *llob6.1* and *llob7.1* on LG6 and LG7, respectively. The QTL *llob6.1* had a major effect on eggplant leaf division, explaining 55.23% of the phenotypic variation with an LOD threshold of 8.99. For *llob7.1*, the physical locations of the three markers were 81.94-87.96 Mb and 38.09 Mb. For LPN, two QTL loci were located on LG7 and LG8, *lpn7.1* and *lpn8.1*, with contribution rates of 9.69 and 7.39%, respectively. The physical region of *lpn7.1* was 17.29-33.55 Mb.

One major-effect QTL (*lpc10.1*) detected for LPC was located on LG10 and explained 36.95% of the phenotypic variation, with an LOD score of 27.65. The QTL for leaf vein color, *lvc10.1*, was also on LG10, and had an overlapped genetic region with *lpc10.1*. The leaf vein color variation explained by *vc10.1* was 52.10%.

DISCUSSION

The Mapping Populations in Eggplant

Eggplant linkage maps were constructed with both intra- and inter-specific populations (Nunome et al., 2001, 2003; Doganlar et al., 2002a; Wu et al., 2009; Barchi et al., 2010, 2012, 2018; Miyatake et al., 2012, 2016; Lebeau et al., 2013; Portis et al., 2014). The intraspecific genetic map generated using an F_2 population derived from 305E40 × 67/3 has wide application. Researchers managed to reuse the F_2 materials by cutting and grafting the established vegetative cuttings and then obtained phenotypic data

TABLE 3 | Detailed information on QTLs detected for the eight traits in the interspecific F_2 population.

No.	QTL	Linkage group ID	SNP no.	Linkage map position (cM)	Interval size (cM)	Genome position (Mb)	Interval size (Mb)	LOD	ADD	DOM	PVE
1	*msh5.1*	E05	11	39.35–44.38	5.03	5.98–7.93 13.57–19.22	1.95 5.65	2.29	3.08	2.15	6.80
2	*fl1.1*	E01	2	65.80–66.01	0.21	28.69–35.85	7.16	3.44	−3.09	−5.84	12.00
3	*fl5.1*	E05	11	4.10–11.96	7.85	0.21–2.71	2.5	2.27	5.39	−0.76	13.06
4	*fl7.1*	E07	5	1.61–9.24	7.63	2.38–3.08	0.7	2.28	2.16	−3.29	4.08
5	*fl12.1*	E12	3	74.02–74.55	0.53	18.05–31.63	13.58	3.77	−4.72	8.94	24.05
6	*fd3.1*	E03	2	44.28–44.43	0.15	4.46–4.97	0.51	2.47	−1.31	−3.66	11.98
7	*fd3.2*	E03	3	69.81–70.40	0.60	60.33–66.04	5.71	2.36	−0.60	−3.38	8.50
8	*fd11.1*	E11	4	71.47–71.87	0.40	58.64–59.84	1.2	2.37	0.63	3.28	7.95
9	*fd11.2*	E11	4	82.29–82.98	0.68	13.15–28.19 55.83–58.83	15.04 0	2.09	0.49	3.04	6.62
10	*fs1.1*	E01	2	87.87–88.58	0.71	110.94–124.29	13.35	3.18	0.09	−0.14	6.92
11	*fs1.2*	E01	4	91.93–92.46	0.54	70.35–72.64	2.29	2.75	0.11	−0.15	8.71
12	*fs1.3*	E01	3	94.23–94.47	0.24	111.30–111.91	0.61	3.16	0.12	−0.16	10.92
13	*fs3.1*	E03	2	44.98–45.32	0.34	73.35–73.74	0.39	2.96	−0.14	−0.08	10.76
14	*llob6.1*	E06	4	71.44–81.97	10.53	104.00–105.00	1	8.99	1.45	−0.15	55.23
15	*llob7.1*	E07	3	73.48–74.35	0.87	81.94–87.96	6.02	2.63	0.61	0.00	9.68
16	*lpn7.1*	E07	2	68.32–69.38	1.06	17.29–33.55	16.26	2.42	0.65	0.24	9.69
17	*lpn8.1*	E08	4	68.58–71.00	2.42	103.00–106.00	3	2.44	−0.60	−0.17	7.39
18	*lpc10.1*	E10	9	79.02–79.94	0.92	87.85–95.97	8.12	27.65	−0.53	1.49	36.95
19	*vc10.1*	E10	8	79.44–80.14	0.70	85.12–95.97	10.85	28.75	−0.68	0.66	52.10

(Barchi et al., 2010, 2012, 2018; Portis et al., 2014; Toppino et al., 2016). This population was used to identify QTLs underlying anthocyanin pigmentation, early yield, fruit-related traits, and resistance to *Fusarium* and *Verticillium*. However, the linkage maps derived from 305E40 × 67/3 contained only ∼400 markers.

Similar to what has been observed in other domesticated crops, artificial selection has caused a dramatic reduction in genetic variation in the cultivated eggplant germplasm, resulting in limited polymorphisms within intraspecific populations and reduced resistance (Daunay et al., 1991; Rotino et al., 2014). Wild Solanum species, which include *S. linnaeanum*, *S. aculeatissimum*, *S. sisymbrifolium*, and *S. torvum*, represent valuable reservoirs of potentially useful resistant alleles for eggplant breeding (Collonnier et al., 2001; Frary et al., 2003; Daunay, 2008; Liu et al., 2015). *S. linnaeanum* is closely related to *S. melongena* (Mace et al., 1999) and was reported to exhibit resistance to Verticillium wilt, black root rot, potato virus, and salinity (Daunay et al., 1991). Fertile F_1 hybrids were only obtained from a cross between *S. linnaeanum* and *S. melongena* (Collonnier et al., 2001). Doganlar et al. (2002a) constructed the first interspecific genetic map with 334 points in eggplant by crossing *S. linnaeanum* MM195 with *S. melongena* MM738. Since then, this interspecific population has been widely used in QTL mapping for fruit, flower, and leaf characteristics, as well as comparative studies with tomato (Doganlar et al., 2002a,b; Frary et al., 2003, 2014; Wu et al., 2009). The genetic map with increased resolution constructed by Frary et al. (2014) reached 735 markers. Nevertheless, the population size of *S. linnaeanum* MM195 × *S. melongena* MM738 was relatively small, containing only 58 individuals.

In the present study, we performed interspecific hybridization between eggplant cultivar "1836" and the wild relative *S. linnaeanum* "1809." An F_1 hybrid was successfully generated and was then self-pollinated to produce an F_2 population with 121 individuals. However, some of the F_2 plants could not bear fruits due to their interspecific nature. The enlarged interspecific population provided raw materials for the introgression of valuable traits from wild species into eggplant cultivars and the identification of QTLs controlling domestication-related traits.

Construction of a Saturated SNP-Based Genetic Map Using SLAF-Seq

To date, over twenty linkage maps have been constructed for eggplant using various types of molecular markers, including RAPDs, AFLPs, and conserved ortholog set (II) (COS/COSII) markers, RFLPs, and SSRs (**Table 4**). Most of them are not saturated due to the low frequency of DNA polymorphism. In the present study, we developed a high-density SNP-based genetic linkage map in eggplant using SLAF sequencing.

The first genetic map in eggplant contained 181 markers, all of which were dominant markers (RAPDs and AFLPs) that were sorted into 21 linkage groups (Nunome et al., 2001). After that, several linkage maps were constructed using mostly dominant markers (Doganlar et al., 2002a,b; Barchi et al., 2010), although the co-dominant SSR markers were introduced into map construction (Nunome et al., 2003). Nunome et al. (2009) developed an enriched SSR-based genetic map containing 236 SSR markers, which were assigned into 14 linkage groups spanning 959.1 cM, with a mean marker interval of 4.3 cM. However, the marker density was still far from saturated. SNPs are the most abundant and stable form of genetic variation in most plant genomes, which have outstanding advantages for the construction of saturated genetic maps. Fukuoka et al. (2012) constructed an integrated linkage map using two mapping populations that include 952 markers (313 SSRs and 623 SNPs) spanning 1285.5 cM. In 2014, another integrated map was constructed with 1745 markers to facilitate eggplant genome assembly, which includes 547 SNPs and 221 SSRs, spanning 1280.6 cM (Hirakawa et al., 2014). Although the marker number was improved in the two integration maps, the applications in QTL mapping were rather limited.

SLAF sequencing takes advantage of high-throughput sequencing and genotyping, providing a powerful tool for genome-wide SNP discovery and marker development. In the present study, we conducted large-scale SNP screening, and 13,455,526 SNP markers were developed, from which 11,386,169 SNPs were successfully encoded. However, a considerable amount of the SNPs were segregation distorted in the interspecific F_2 population. Thus, we performed a strict marker filtering process before map construction. Finally, a saturated genetic map with 2,122 high-quality SNP markers was constructed using SLAF-seq (**Table 2** and **Figure 2**), which is a significant improvement in marker number as compared to the individual maps (Frary et al., 2014; Miyatake et al., 2016; Barchi et al., 2018) as well as the two integrated linkage maps (Fukuoka et al., 2012; Hirakawa et al., 2014). The total genetic length of the SNP-based linkage map was 1530.75 cM, and the average marker distance was narrowed down to 0.72 cM. The marker number in each LG ranged from 90 to 273 SNPs, with mean marker intervals ranging from 0.43 cM to 1.4 cM. This high-density genetic map establishes a foundation for accurate and reliable mapping of QTLs, as well as the identification of candidate genes underlying important traits in eggplant.

QTL Mapping of Morphological Traits in Eggplant

Eggplants exhibit wide biodiversity among local landraces and wild relatives, with considerable variations in fruit size and color, leaf morphology, and pathogen resistance. Unlike most of the other major Solanaceous crops, which are native to the New World, eggplant has a unique phylogeny of Old World domestication that occurred in India and Southern China (Fukuoka et al., 2010; Meyer et al., 2012; Cericola et al., 2013; Albert and Chang, 2014). The wild forms of eggplant are usually pricky with small, bitter fruits; however, selection during domestication resulted in elongated and palatable fruits with fewer prickles in cultivated eggplant (Choudhury, 1995). The two parental lines used in the present study, *S. melongena* "1836" and the wild *S. linnaeanum* "1809," have contrasting phenotypes (**Figure 1**), making them valuable for investigating the molecular mechanisms underlying domestication-related traits.

TABLE 4 | Summery of previously reported genetic maps in eggplant.

Cross parent	Interspecific or intraspecific	Group type	Population size	Total Linkage group no.	genetic distance	Marker no.	Marker type	References
EPL-1 × WCGR112-8	Intraspecific	F_2	168	21	779.2	181	RAPD, AFLP	Nunome et al., 2001
Solanum linnaeanum MM195 × *S. melongena* MM738	Interspecific	F_2	58	12	1480	334	tomato cDNA, genomic DNA, COS, RFLP	Doganlar et al., 2002a
S. linnaeanum MM195 × *S. melongena* MM738	Interspecific	F_2	58	12		207	RFLP	Doganlar et al., 2002b
S. linnaeanum MM195 × *S.melongena* MM738	Interspecific	F_2	58	12		207	RFLP	Frary et al., 2003
EPL-1 × WCGR112-8	Intraspecific	F_2	168	17	716.9	162	RAPD, AFLP, SSR	Nunome et al., 2003
S. sodomeum(= *S. linneanum*) × *S. melongena* Buia	Interspecific	F_2	48	13	736	273	RAPD, AFLP	Sunseri et al., 2003
EPL-1 × WCGR112-8	Intraspecific	F_2	94	14	959.1	236	SSR	Nunome et al., 2009
S. linnaeanum MM195 × *S. melongena* MM738	Interspecific	F_2	58	12	1535	347	COSII	Wu et al., 2009
305E40 × 67/3	Intraspecific	F_2	141	12	718.7	238	AFLP, SSR, RFLP, Rfo-sa1 CAPS	Barchi et al., 2010
305E40 × 67/3	Intraspecific	F_2	156	12	1389.7	415	SNP, SSR, COS	Barchi et al., 2012
LS1934 × WCGR112-8, AE-P03 × LS1934	Intraspecific	F_2	90, 93	12	1285.5	952	SSR, SNP	Fukuoka et al., 2012
AE-P03 × LS1934	Intraspecific	F_2	135	12	1414.6	250	SSR, SNP SSR, SNP	Miyatake et al., 2012
Nakate-Shinkuro × AE-P03	Intraspecific	F_2	93	12	1153.8	174		
MM738 × AG91-25	Intraspecific	F_6	178	18	884	119	AFLP, SSR, SRAP	Lebeau et al., 2013
S. linnaeanum MM195 × *S. melongena* MM738	Interspecific	F_2	58	12		736	AFLP, RFLP, COSII	Frary et al., 2014
305E40 × 67/3	Intraspecific	F_2	156	12	1389.7	415	SNP, SSR, COS	Portis et al., 2014
LS1934 × WCGR112-8, EPL-1 × WCGR112-8	Intraspecific	F_2	90, 120	12	1280.6	1745	SNP, SSR	Hirakawa et al., 2014
LS1934 × WCGR112-8	Intraspecific	F_2	90	12	1280.6	1193	SNP,	Miyatake et al., 2016
EPL-1 × WCGR112-8	Intraspecific	F_2	120	12	1280.6	602	SSR	
AE-P03 × LS1934	Intraspecific	F_2	93	12	1285.5	952		
305E40 × 67/3	Intraspecific	F_2	156	12	1389.7	415	SNP, SSR, COS	Toppino et al., 2016
305E40 × 67/3	Intraspecific	F_2	156	12	1390	418	SNP, SSR, COS, HRM	Barchi et al., 2018

Using the high-density SNP map and the interspecific F_2 population, we identified a total of 19 QTLs for main stem length and fruit and leaf morphology (**Table 3** and **Figure 2**). While no QTL loci were detected on three LGs (2, 4, and 9), all of the other nine LGs had QTL distributions. The phenotypic variance explained by the QTLs ranged between 4.08 and 55.23%, and the genetic distance interval varied from 0.15 to 10.53 cM. We detected one QTL for main stem height (*msh5.1*) on LG5, explaining 6.8% of the phenotypic variation. The genetic interval was 5.03 cM, covering 11 SNPs. Previous reports on QTLs for eggplant fruit and leaf traits are rather limited; we summarize previously mapped QTLs related to the eight traits in the present study in **Supplementary Table S3**. Doganlar et al. (2002b) identified three QTLs for eggplant fruit length using the F_2 population derived from *S. linnaeanum* "MM195" × *S. melongena* "MM738" and RFLP markers; the three QTLs (i.e., *fl2.1*, *fl9.1*, and *fl11.1*) accounted for 23-29% of the fruit length variation. Using the same population and a 736-point genetic map, Frary et al. (2014) detected five QTLs impacting fruit length that were distributed on LG2, 7, and 9 (*fl1.1*, *fl2.1*, *fl2.2*, *fl7.1*, and *fl9.1*). Another report used an intraspecific population derived from 305E40 × 67/3 and a genetic map with 415 markers, and six QTLs affecting fruit length were detected over six LGs: LG1, 2, 3, 7, 8, and 11 (Portis et al., 2014). In the present study, four QTLs (*fl1.1*, *fl5.1*, *fl7.1*, and *fl12.1*) were identified for fruit length, and the QTL *fl12.1* explained 24.05% of the variation for fruit length. Thus, fruit length-related QTLs in eggplant are distributed on nine different LGs.

For fruit diameter, we detected four QTLs on LG3 (*fd3.1*, *fd3.2*) and LG11 (*fd11.1, fd11.2*), with the genetic interval of each QTL ranging between 0.15 and 0.68 cM. These QTLs accounted for 6.62 to 11.98% of the observed phenotypic variation. Doganlar et al. (2002b) identified two QTLs on LG1 (*fd1.1*) and LG11 (*fd11.1*), which explained 17% of the total variation for FD. Whereas Portis et al. (2014) adopted three FD parameters for the intraspecific population (i.e., fd1/2, fd3/4, and fdmax), three to seven QTLs were mapped on LG2, 3, 4, 7, 11, and 12. FS-related QTLs were also detected in the two aforementioned studies, which were distributed on LG1, 2, 3, 7, and 11. In the present study, we identify four QTLs for fruit shape index, among which three were on LG1 (*fs1.1*, *fs1.2*, *fs1.3*) and one on LG3 (*fs3.1*). The

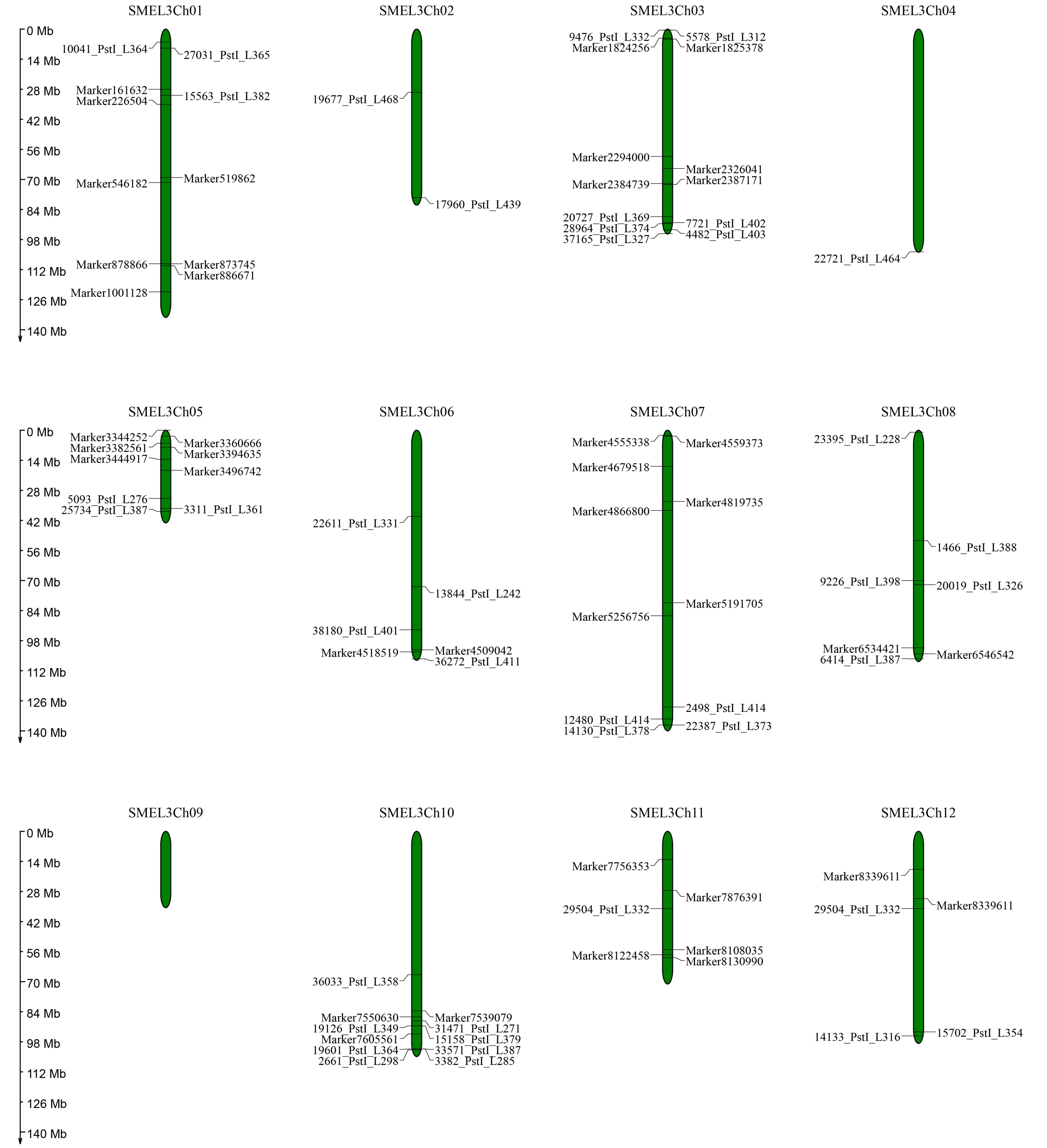


FIGURE 3 | Graphic view of the distribution of markers associated with related traits on eggplant chromosomes.

genetic locations of *fs1.1*, *fs1.2*, and *fs1.3* were very close together, suggesting that they may function as a single locus. Collectively, the three QTLs explained 26.55% of the fruit shape variation.

In total, six QTLs were detected for leaf morphology-related traits in this study. LPC and LVC were highly correlated (**Supplementary Table S2**), and as expected, *vc10.1* had an overlapping genetic location with *lpc10.1*. *llob6.1* was identified as a major-effect QTL that accounted for 55.23% of the phenotypic variation. In previous studies, Frary et al. (2003) identified two QTLs for leaf lobing, on LG6 (*llob6.1*) and LG10 (*llob10.1*); after increasing the marker density on the original map, Frary et al. (2014) identified four QTLs on LG5, 6, and 7 using the same population. In the previous studies, each QTL interval was only covered by one or two markers (**Supplementary Table S3**); with the 2,122-point SNP-based map, up to 11 SNPs were harbored in a single QTL locus in this study. This increase in mapped markers could better facilitate the fine mapping of these QTLs in further analysis.

To better demonstrate the mechanisms underlying fruit and leaf morphology traits in eggplant, we performed comparative analysis between the QTLs in the present study and previous QTL analysis and association studies (Barchi et al., 2012; Cericola et al., 2014; Portis et al., 2014, 2015) based on marker sequences and the eggplant reference genome (Barchi et al., 2019). A graphic view of the distribution of the QTLs and markers associated with related traits on eggplant chromosomes was produced (**Figure 3** and **Supplementary Table S3**). There were no relevant markers or QTL loci on chromosome 9, whereas all of the other 11 chromosomes had marker distribution. The SNPs in the QTLs we located could be associated with some of the markers in previous studies. For example, markers 19126_PstI_L349, 31471_PstI_L271, and 15158_PstI_L379 were shown to be related to the anthocyanin content of leaf veins (Portis et al., 2014), and these markers are also located in *lpc10.1* and *vc10.1* in the present study. Thus, we speculated that there were candidate genes related to anthocyanin accumulation in leaves in this region. Marker 29504_PstI_L332, which has been related to fruit length, diameter, and fruit shape (Portis et al., 2014), is close to *fl12.1* in the present study. Nonetheless, there are also inconsistencies among different studies. The QTL controlling fruit diameter determined by markers 9476_PstI_L332 and 5578_PstI_L312 was anchored to the eggplant chromosome 3, close to the QTL *fd3.1* we detected. In addition, marker 36272_PstI_L411 was related to the leaf prickle number trait but close to the QTL of leaf lobing we located. Notably, some markers are relatively far apart on the chromosome in terms of physical location, whereas they were close on the genetic map. This is likely due to errors in either linkage distance calculation or misassembly of the eggplant genome. Another possible reason is that the two parents of the mapping population were different from the genome sequencing eggplant material, especially the wild species *S. linnaeanum*, resulting in additional genetic polymorphism.

CONCLUSION

In conclusion, the high-density SNP-based genetic map and QTLs controlling agronomic traits for eggplant in the present study provide an important foundation for developing tightly linked markers for marker-assisted breeding, as well as fine mapping and gene mining of related traits, especially the QTLs presented in the interspecific population, which could facilitate the demonstration of eggplant domestication. We also assigned the QTLs to eggplant chromosomes and have provided the physical positions of the markers and their sequences. QTL loci of the same traits in multiple studies should be anchored to the same high-quality eggplant genome; the hotspots controlling those traits could then be determined based on repeatability, and further precise predictions could be made of the candidate genes for functional analysis. However, more marker sequence information corresponding to the QTLs needs to be disclosed.

DATA AVAILABILITY STATEMENT

Raw sequence reads have been submitted to the NCBI Sequence Read Archive under the accession number PRJNA577305. The datasets supporting the conclusions drawn in this study are included within the manuscript and the **Supplementary Tables**.

AUTHOR CONTRIBUTIONS

QW conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the manuscript, and approved the final draft. WW analyzed the data, authored or reviewed drafts of the manuscript, and approved the final draft. TH, HH, and JW analyzed the data, contributed reagents, materials, and analysis tools, authored or reviewed drafts of the manuscript, and approved the final draft. CB conceived and designed the experiments, authored or reviewed drafts of the manuscript, and approved the final draft.

FUNDING

The present study was supported by the Natural Science Foundation of Zhejiang Province (grant number LQ18C150004), Major Science and Technology Projects of Zhejiang (grant number 2016C02051-2-1), and Youth Program of Zhejiang Academy of Agricultural Sciences (grant number 2019R23R08E03).

ACKNOWLEDGMENTS

We would like to thank Dr. Zheng Zheng and Dr. Ziqi Sun of Henan Academy of Agricultural Sciences for their help in genetic map construction using Joinmap v5.0 software.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00178/full#supplementary-material

FIGURE S1 | Distribution of the QTLs on eggplant chromosomes.

TABLE S1 | Genotyping results of the markers and detailed information of the SNP-based genetic linkage map of eggplant.

TABLE S2 | Phenotypic data and ANOVA analysis of the interspecific F_2 population.

TABLE S3 | Summary of the previously reported QTLs related to fruit and leaf traits of eggplant.

REFERENCES

Albert, V. A., and Chang, T. H. (2014). Evolution of a hot genome. *PNAS* 111, 5069–5070.

Ali, I., Teng, Z., Bai, Y., Yang, Q., Hao, Y., Hou, J., et al. (2018). A high density SLAF-SNP genetic map and QTL detection for fibre quality traits in *Gossypium hirsutum*. *BMC Genomics* 19:879. doi: 10.1186/s12864-018-5294-5

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376. doi: 10.1371/journal.pone.0003376

Barchi, L., Lanteri, S., Portis, E., Stàgel, A., Valè, G., Toppino, L., et al. (2010). Segregation distortion and linkage analysis in eggplant (*Solanum melongena* L.). *Genome* 53, 805–815. doi: 10.1139/g10-073

Barchi, L., Lanteri, S., Portis, E., Valè, G., Volante, A., Pulcini, L., et al. (2012). A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS One* 7:e43740. doi: 10.1371/journal.pone.0043740

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., et al. (2019). A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Sci. Rep.* 9:11769. doi: 10.1038/s41598-019-47985-w

Barchi, L., Toppino, L., Valentino, D., Bassolion, L., Portis, E., Lanteri, S., et al. (2018). QTL analysis reveals new eggplant loci involved in resistance to fungal wilts. *Euphytica* 214:20.

Cericola, F., Portis, E., Lanteri, S., Toppino, L., Barchi, L., Acciarri, N., et al. (2014). Linkage disequilibrium and genome-wide association analysis for anthocyanin pigmentation and fruit color in eggplant. *BMC Genomics* 15:896. doi: 10.1186/1471-2164-15-896

Cericola, F., Portis, E., Toppino, L., Barchi, L., Acciarri, N., Ciriaci, T., et al. (2013). The population structure and diversity of eggplant from asia and the mediterranean basin. *PLoS One* 8:e73702. doi: 10.1371/journal.pone.0073702

Choudhury, B. (1995). *Evolution in Crop Plants, edited by J. Smartt and N. W. Simmonds*. New York, NY: John Wiley & Sons, 464–465.

Collonnier, C., Fock, I., Kashyap, V., Rotino, G. L., Daunay, M. C., Lian, Y., et al. (2001). Applications of biotechnology in eggplant. *Plant Cell Tissue Organ Cult.* 65, 91–107.

Daunay, M. C. (2008). "Eggplant," in *Handbook of Crop Breeding., Vegetables: Fabaceae, Liliaceae, Umbelliferae, and Solanaceae*, eds J. Prohens, and F. Nuez (New York, NY: Springer), 163–220.

Daunay, M. C., Fary, A., and Doganlar, S. (2001). "Genetic resources of eggplant (*Solanum melongena* L.) and allied species: a new challenge for molecular geneticists and eggplant breeders," in *Solanaceae V. Advances in Taxonomy and Utilization*, eds R. G. van den Berg, G. W. M. Barendse, G. M. van der Weerden, and C. Mariani (Nijmegen: Nijmegen University Press), 251–274.

Daunay, M. C., Lester, R. N., and Laterrot, H. (1991). *The Use of Wild Species for the Genetic Improvement of Brinjal Eggplant Solanum melongena and Tomato Lycopersicon Esculentum. Solanaceae III: Taxonomy, Chemistry, Evolution*, Vol. 27. Richmond: Royal Botanic Gardens Kew, 389–413.

Doganlar, S., Frary, A., Daunay, M. C., Lester, R. N., and Tanksley, S. D. (2002a). A comparative genetic linkage map of eggplant (*Solanum melongena* L.) and its implications for genome evolution in the Solanaceae. *Genetics* 161, 1697–1711.

Doganlar, S., Frary, A., Daunay, M. C., Lester, R. N., and Tanksley, S. D. (2002b). Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in Eggplant. *Genetics* 161, 1713–1726.

Frary, A., Doganlar, S., Daunay, M. C., and Tanksley, S. D. (2003). QTL analysis of morphological traits in eggplant and implications for conservation of gene function during evolution of solanaceous species. *Theor. Appl. Genet.* 107, 359–370. doi: 10.1007/s00122-003-1257-5

Frary, A., Frary, A., Daunay, M.-C., Huvenaars, K., Mank, R., and Doğanlar, S. (2014). QTL hotspots in eggplant (Solanum melongena) detected with a high resolution map and CIM analysis. *Euphytica* 197, 211–228. doi: 10.1007/s10681-013-1060-6

Fukuoka, H., Miyatake, K., Nunome, T., Negoro, S., and Ohyama, A. (2012). Development of gene-based markers and construction of an integrated linkage map in eggplant by using *Solanum orthologous* (SOL) gene sets. *Theor. Appl. Genet.* 125, 47–56. doi: 10.1007/s00122-012-1815-9

Fukuoka, H., Yamaguchi, H., Nunome, T., Negoro, S., Miyatake, K., and Ohyama, A. (2010). Accumulation, functional annotation, and comparative analysis of expressed sequence tags in eggplant (*Solanum melongena* L.), the third pole of the genus *Solanum* species after tomato and potato. *Gene* 450, 76–84. doi: 10.1016/j.gene.2009.10.006

He, Y. X., Yuan, W. J., Dong, M. F., Han, Y. J., and Shang, F. D. (2017). The first genetic map in sweet osmanthus (*Osmanthus fragrans* lour.) using specific locus amplified fragment sequencing. *Front. Plant Sci.* 8:1621. doi: 10.3389/fpls.2017.01621

Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., and Fukuoka, H. (2014). Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the old world. *DNA Res.* 21, 649–660. doi: 10.1093/dnares/dsu027

Hu, X. H., Zhang, S. Z., and Miao, H. R. (2018). High-density genetic map construction and identification of qtls controlling oleic and linoleic acid in peanut using SLAF-seq and SSRs. *Sci. Rep.* 8:5479. doi: 10.1038/s41598-018-23873-7

Huang, S., Ding, D., Tang, W., and Liu, Y. (2013). Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* 4:2640. doi: 10.1038/ncomms3640

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39. doi: 10.1038/ng.1018

Lebeau, A., Gouy, M., Daunay, M. C., Wicker, E., Chiroleu, F., Prior, P., et al. (2013). Genetic mapping of a major dominant gene for resistance to *Ralstonia solanacearum* in eggplant. *Theor. Appl. Genet.* 126, 143–158. doi: 10.1007/s00122-012-1969-5

Liu, D., Ma, C., Hong, W., Huang, L., Liu, M., Liu, H., et al. (2014). Construction and analysis of high-density linkage map using high-throughput sequencing data. *PLoS One* 9:e98855. doi: 10.1371/journal.pone.0098855

Liu, J., Zheng, Z., Zhou, X., Feng, C., and Zhuang, Y. (2015). Improving the resistance of eggplant (*Solanum melongena* L.) to Verticillium wilt using wild species *Solanum linnaeanum*. *Euphytica* 201, 463–469. doi: 10.1007/s10681-014-1234-x

Mace, E. S., Lester, R. N., and Gebhardt, C. G. (1999). AFLP analysis of genetic relationships among the cultivated eggplant, *Solanum melongena* L., and wild relatives (Solanaceae). *Theor. Appl. Genet.* 99, 626–633. doi: 10.1007/s001220051277

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meyer, R. S., Karol, K. G., Little, D. P., Nee, M. H., and Litt, A. (2012). Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. *Mol. Phylogenet. Evol.* 63, 685–701. doi: 10.1016/j.ympev.2012.02.006

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17, 240–248. doi: 10.1101/gr.5681207

Miyatake, K., Saito, T., Negoro, S., Yamaguchi, H., Nunome, T., Ohyama, A., et al. (2016). Detailed mapping of a resistance locus against Fusarium wilt in cultivated eggplant (*Solanum melongena* L.). *Theor. Appl. Genet.* 129, 357–367. doi: 10.1007/s00122-015-2632-8

Miyatake, K., Saito, T., Negoro, S., Yamaguchi, H. N., Unome, T., Ohyama, A., et al. (2012). Development of selective markers linked to a major QTL for parthenocarpy in eggplant (*Solanum melongena*, L.). *Theor. Appl. Genet.* 124, 1403–1413. doi: 10.1007/s00122-012-1796-8

Murray, M., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4326. doi: 10.1093/nar/8.19.4321

Nunome, T., Ishiguro, K., Yoshida, T., and Hirai, M. (2001). Mapping of fruit shape and color development traits in eggplant (*Solanum melongena* L.) based on RAPD and AFLP markers. *Breed. Sci.* 51, 9–26.

Nunome, T., Negoro, S., Kono, I., Kanamori, H., Miyatake, K., Yamaguchi, H., et al. (2009). Development of SSR markers derived from SSR-enriched genomic library of eggplant (*Solanum melongena* L.). *Theor. Appl. Genet.* 119, 1143–1153. doi: 10.1007/s00122-009-1116-0

Nunome, T., Suwabe, K., Iketani, H., Hirai, M., and Wricke, G. (2003). Identification and characterization of microsatellites in eggplant. *Plant Breed.* 122, 256–262. doi: 10.3389/fpls.2018.00401

Ogden, R. (2011). Unlocking the potential of genomic technologies for wildlife forensics. *Mol. Ecol. Resour.* 11, 109–116. doi: 10.1111/j.1755-0998.2010.02954.x

Portis, E., Barchi, L., Toppino, L., Sergio, L., Nazzareno, A., Nazzaremo, F., et al. (2014). QTL mapping in eggplant reveals clusters of yield-related loci and orthology with the tomato genome. *PLoS One* 9:e89499. doi: 10.1371/journal.pone.0089499

Portis, E., Cericola, F., Barchi, L., Toppino, L., Acciarri, N., Pulcini, L., et al. (2015). Association mapping for fruit, plant and leaf morphology traits in eggplant. *PLoS One* 10:e0135200. doi: 10.1371/journal.pone.0135200

Qi, Z., Huang, L., Zhu, R., Xin, D., Liu, C., Han, H., et al. (2014). A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS One* 9:e104871. doi: 10.1371/journal.pone.0104871

Rotino, G. L., Sala, T., and Toppino, L. (2014). "Eggplant. Book," in *Alien Gene Transfer in Crop Plants, vol 2, 381, Achievements and Impacts*, Chap. 16, eds A. Pratap, and J. Kumar (New York, NY: Springer).

Sun, X. W., Liu, D. Y., Zhang, X. F., Li, W. B., Liu, H., Hong, W. G., et al. (2013). SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8:e58700. doi: 10.1371/journal.pone.0058700

Sunseri, F., Sciancalepore, A., Martelli, G., Acciarri, N., Rotino, G. L., Valentino, D., et al. (2003). Development of RAPD-AFLP map of eggplant and improvement of tolerance to *Verticillium* wilt. *Acta Hortic.* 625, 107–110.

Toppino, L., Barchi, L., Lo Scalzo, R., Palazzolo, E., Francese, G., Fibiani, M., et al. (2016). Mapping quantitative trait loci affecting biochemical and morphological fruit properties in eggplant (*Solanum melongena* L.). *Front. Plant Sci.* 4:256. doi: 10.3389/fpls.2016.00256

van Os, H., Stam, P., Visser, R. G. F., and Eck, H. J. V. (2005). SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. *Theor. Appl. Genet.* 112, 187–194. doi: 10.1007/s00122-005-0124-y

Wang, S., Meyer, E., McKay, J. K., and Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Meth.* 9, 808–810. doi: 10.1038/nmeth.2023

Wei, Q., Wang, Y., Qin, X., Zhang, Y., Zhang, Z., Wang, J., et al. (2014). An SNP-based saturated genetic map and QTL analysis of fruit-related traits in cucumber using specific-length amplified fragment (SLAF) sequencing. *BMC Genomics* 15:1158. doi: 10.1186/1471-2164-15-1158

Wu, F., Eannetta, N. T., Xu, Y., and Tanksley, S. D. (2009). A detailed synteny map of the eggplant genome based on conserved ortholog set II (COSII) markers. *Theor. Appl. Genet.* 118, 927–935. doi: 10.1007/s00122-008-0950-9

Zhang, Y., Li, W., Lin, Y., Zhang, L., Wang, C., and Xu, R. (2018). Construction of a high-density genetic map and mapping of QTLs for soybean (Glycine max) agronomic and seed quality traits by specific length amplified fragment sequencing. *BMC Genomics* 19:641.

Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., et al. (2013). Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biol.* 13:141. doi: 10.1186/1471-2229-13-141

Zhang, Z., Shang, H., Shi, Y., Huang, L., Li, J., Ge, Q., et al. (2016). Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to Quantitative Trait Loci (QTL) analysis for boll weight in upland cotton (*Gossypium hirsutum.*). *BMC Plant Biol.* 16:79. doi: 10.1186/s12870-016-0741-4

Zhu, W. Y., Huang, L., Chen, L., Yang, J. T., Wu, J. N., Qu, M. L., et al. (2016). A high-density genetic linkage map for cucumber (*Cucumis sativus* L.): based on specific length amplified fragment (SLAF) sequencing and QTL analysis of fruit traits in cucumber. *Front. Plant Sci.* 7:437. doi: 10.3389/fpls.2016.00437

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fgene.2020.00222

Comparison of SNP Calling Pipelines and NGS Platforms to Predict the Genomic Regions Harboring Candidate Genes for Nodulation in Cultivated Peanut

Ze Peng[1†], *Zifan Zhao*[1†], *Josh Paul Clevenger*[2], *Ye Chu*[3], *Dev Paudel*[1], *Peggy Ozias-Akins*[3] *and Jianping Wang*[1,4*]

[1] Agronomy Department, University of Florida, Gainesville, FL, United States, [2] Center for Applied Genetic Technologies, University of Georgia, Athens, GA, United States, [3] Genetic and Genomics and Department of Horticulture, Institute of Plant Breeding, University of Georgia, Tifton, Georgia, [4] Genetics Institute and Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, United States

OPEN ACCESS

Edited by:
Wenqin Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Chen Yuning,
Chinese Academy of Agricultural Sciences, China
Qi Li,
Shanghai Institutes for Biological Sciences (CAS), China

***Correspondence:**
Jianping Wang
wangjp@ufl.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 18 December 2019
Accepted: 25 February 2020
Published: 24 March 2020

Citation:
Peng Z, Zhao Z, Clevenger JP, Chu Y, Paudel D, Ozias-Akins P and Wang J (2020) Comparison of SNP Calling Pipelines and NGS Platforms to Predict the Genomic Regions Harboring Candidate Genes for Nodulation in Cultivated Peanut. Front. Genet. 11:222. doi: 10.3389/fgene.2020.00222

Cultivated peanut (*Arachis hypogaea* L.) forms root nodules to enable a symbiotic relationship with rhizobia for biological nitrogen fixation. To understand the genetic factors of peanut nodulation, it is fundamental to genetically map and clone the genes involved in nodulation. For genetic mapping, high throughput genotyping with a large number of polymorphic markers is critical. In this study, two sets of sister recombinant inbred lines (RILs), each containing a nodulating (Nod+) and non-nodulating (Nod-) line, and their Nod+ parental lines were extensively genotyped. Several next generation sequencing (NGS) methods including target enrichment sequencing (TES), RNA-sequencing (RNA-seq), genotyping by sequencing (GBS), and the 48K Axiom *Arachis2* SNP array, and various analysis pipelines were applied to identify single nucleotide polymorphisms (SNP) among the two sets of RILs and their parents. TES revealed the largest number of homozygous SNPs (15,947) between the original parental lines, followed by the Axiom *Arachis2* SNP array (1,887), RNA-seq (1,633), and GBS (312). Among the five SNP analysis pipelines applied, the alignment to A/B genome followed by HAPLOSWEEP revealed the largest number of homozygous SNPs and highest concordance rate (79%) with the array. A total of 222 and 1,200 homozygous SNPs were polymorphic between the Nod+ and Nod− sister RILs and between their parents, respectively. A graphical genotype map of the sister RILs was constructed with these SNPs, which demonstrated the candidate genomic regions harboring genes controlling nodulation across the whole genome. Results of this study mainly provide the pros and cons of NGS and SNP genotyping platforms for genetic mapping in peanut, and also provide potential genetic resources to narrow down the genomic regions controlling peanut nodulation, which would lay the foundation for gene cloning and improvement of nitrogen fixation in peanut.

Keywords: genotyping by sequencing, nodulation, peanut, RNA sequencing, single nucleotide polymorphism, SNP array, target enrichment sequencing

INTRODUCTION

Peanut (*Arachis hypogaea* L.) is one of the most important oilseed crops grown worldwide. As a legume species, peanut can form a symbiotic relationship with rhizobia to biologically fix nitrogen, thus reducing the amount of synthetic nitrogen fertilizers applied in the growing season. The symbiotic process and molecular mechanisms have been extensively studied in two model legume species *Lotus japonicas* and *Medicago truncatula,* where rhizobia enter into the host plant via an intracellular root hair pathway (Oldroyd, 2013). Many genes have been characterized in the symbiotic pathway and some rhizobial small RNA fragments were also reported to play a regulatory role (Ren et al., 2019). In peanut, rhizobia infect plants via the intercellular crack entry, which is less studied and not well understood (Peng et al., 2017a). Non-nodulating (Nod-) peanut plants, first reported by Gorbet and Burton (1979), are important materials for dissecting the genetic factors of peanut nodulation. The Nod- peanut plants were first identified in an F_3 population from a cross between two nodulating (Nod+) genotypes 487A-4-1-2 and PI 262090 (Gorbet and Burton, 1979). Several gene inheritance models were subsequently proposed by investigating segregation ratios in populations segregating for nodulation, including the two-gene (Nigam et al., 1980), and three-gene (Dutta and Reddy, 1988; Gallo-Meagher et al., 2001) models. However, no nodulation genes have been either identified or characterized. A transcriptome study using root samples from two sets of recombinant inbred lines (RILs) with Nod+ and Nod− phenotype revealed hundreds of differentially expressed genes (DEGs) upon infection with rhizobia (Peng et al., 2017a). In addition, the same materials were morphologically and genetically characterized to initiate studies on peanut nodulation genes (Peng et al., 2018). A total of 188 simple sequence repeat (SSR) markers were used for genetic characterization, and only a few polymorphic SSRs were obtained between the RILs due to their high genetic similarity. The graphical genotype maps of the RILs were subsequently constructed showing candidate genomic regions controlling peanut nodulation and a total of 22 chromosome regions potentially related with nodulation were revealed between two sets of RILs. However, with a limited number of markers, the maps had a low resolution, which is hard for further fine mapping. With the aid of next generation sequencing (NGS) technologies, the map density could be further improved.

Peanut is an allotetraploid ($2n$ = 2x = 40; AABB; ∼2.7 Gb) with two sub-genomes, A and B, derived from *A. duranensis* and *A. ipaensis*, respectively (Bertioli et al., 2016). The available reference genomes of the two diploid ancestors have made whole-genome resequencing (WGRS) an applicable approach for high throughput genotyping, which was used for genotyping a bi-parental population for high-density genetic map construction and candidate disease resistance gene identification in peanut (Agarwal et al., 2018). Each sample was sequenced at 2∼5 × coverage. However, considering the large genome size and high content of repetitive sequences in the peanut genome, WGRS may still not be the most cost-effective strategy to detect genetic variations, as the per sample cost is still high especially if high coverage is expected (Schwarze et al., 2018). Alternatively, other NGS enabled genotyping methods with reduced genome complexity can be cost-efficient for high throughput genotyping, such as RNA-sequencing (RNA-seq) (Clevenger et al., 2015; Chopra et al., 2016), genotyping by sequencing (GBS) (Tseng et al., 2016), and target-enrichment sequencing (TES) (Peng et al., 2017b), which discover genetic variations from a representative proportion of the genome. In addition, the Axiom *Arachis2* array with 47,837 SNPs can be a cost-efficient and simple method for high throughput genotyping (Clevenger et al., 2018), though it is limited to known single nucleotide polymorphisms (SNPs) only.

As the A and B genomes of peanut are highly similar with a median identity of 93.11% (Bertioli et al., 2016), it has been a big challenge to identify allelic SNPs due to the confounding effect of homoeologous SNPs between the two sub-genomes (Clevenger et al., 2015). Multiple strategies and tools have been developed to resolve this issue. One option to reduce the amount of homoeologous SNPs is to exclusively utilize uniquely mapped reads for subsequent SNP calling (Zhou et al., 2014; Peng et al., 2017b), which led to a decreased number of useful SNPs identified. Alternatively, several other methods have been developed that could use overall mapped reads for SNP calling and filter out homoeologous SNPs afterward. For example, SWEEP (Clevenger and Ozias-Akins, 2015), which utilizes homoeologous SNPs as an anchor to differentiate allelic SNPs, had been successfully applied in peanut (Clevenger et al., 2017; Pandey et al., 2017) with a validation rate of 85% through Sanger sequencing and above 95% through simulation data (Clevenger and Ozias-Akins, 2015). In addition, a machine-learning tool called SNP-ML was also developed to predict allelic SNPs with a validation rate of 75–98% (Korani et al., 2019). An improved version of SWEEP, named HAPLOSWEEP, was developed, which applies a haplotype-based method to identify allelic polymorphisms between genotypes (Clevenger et al., 2018), and it had a validation rate of 74% through genotyping by the Axiom *Arachis2* array. With these methods and tools available for the peanut community, currently no study has been performed to compare these SNP calling and filtering methods, or to compare the effects of mapping reads to the concatenated A + B genome or to A and B genomes separately (A/B).

In this study, to explore the genetic factors and genetic regions controlling nodulation in peanut, SNPs were identified between the two original Nod+ parental lines as well as between two sets of RILs. Three NGS approaches, including TES, RNA-seq, and GBS were applied and compared for SNP identification. To summarize and compare different SNP analysis methods, we have applied and compared two alignment methods (to A + B genome or to A/B genome) and various SNP calling and filtering pipelines using the sequencing data. In addition, the Axiom *Arachis2* array was also used for genotyping and served as a SNP cross-validation platform for identified SNPs. This is the first study to compare different SNP calling and filtering pipelines for various NGS data sources in peanut. Results and suggestions from this study provide insights into SNP identification and genotyping in peanut. The polymorphic genomic regions between the sister

RILs revealed candidate genes controlling peanut nodulation, which will be beneficial for future genetic mapping studies.

MATERIALS AND METHODS

Plant Materials

Two sets of RILs, E4 (Nod−) & E5 (Nod+), and E6 (Nod+) & E7 (Nod−), as well as their parental lines, PI 262090 (Nod+) and UF 487A (Nod+) were included in this study. The pedigree information of these six lines was introduced previously (Peng et al., 2017a). In brief, the two sets of RILs can be traced to two different F_6 lines, which were deived from the cross between PI 262090 and UF 487A. They are also parental lines for two F_2 mapping populations (E4 × E5 and E6 × E7) for genetic mapping of nodulation genes. The morphological and genetic characterizations of the RILs were previously described (Peng et al., 2018). The genomic DNA of the six genotypes was extracted by using the CTAB method (Rogers and Bendich, 1994). DNA concentration and quality were checked using agarose gel and NanoDrop.

Probe Design, Evaluation, and Selection for Target Enrichment Sequencing

To preferably target peanut genes potentially related to nodulation, a series of genes were included for probe design. Firstly, the putative orthologous nodulation-related genes and differentially expressed genes (DEGs) upon infection of rhizobia from the previous report (Peng et al., 2017a) were included (referred to as Class I genes). For these peanut genes, the gene sequences together with 2 Kb upstream and 1 Kb downstream sequences were subjected to probe design. For the Class I genes, if there were more than four peanut genes in the same orthologous group with the nodulation-related gene in model legumes, only the top four genes (based on Blast score) were included for subsequent probe selection. Secondly, for the remaining genes that were annotated in the peanut diploid ancestors' genomes (referred to as Class II genes), only the gene coding sequences were utilized for probe design. The probes were 120 bp long and had no overlap with each other. A total of 3,982 Class I genes were obtained from the previous transcriptome study (Peng et al., 2017a). The sequences of those genes together with the remaining 74,753 Class II gene models in the diploid ancestors' genomes of peanut were submitted for probe design.

A probe could capture or hybridize with the DNA fragments if they share sequence similarity with each other. The genomic regions sharing sequence similarities with probes were considered as probe target regions. However, the capture efficiency would be different for target regions with different similarities. Thus, the number of target regions was investigated for the probes under different alignment identity cutoffs when they were mapped to the genome. The uniqueness and distribution of the designed probes were further evaluated.

To evaluate uniqueness of the designed probes in the genome, the probe sequences were mapped back to the diploid genomes of peanut (A + B) using Blat (Kent, 2002). A hit was defined under cutoff: e-value ≤ 1e-05; alignment identity = alignment length × percentage of identity ≥96 (120 bp × 80% = 96 bp). For easier downstream data analysis, primarily single-hit probes were selected for synthesis. A unique set of single-hit probes was obtained by using CD-HIT-EST (-c 0.8 -aL 0.8 -AL 24 -aS 0.8 -AS 24 -n 5 -T 0 -r 1) (Fu et al., 2012). All single-hit probes covering Class I genes and resistance genes annotated in the genome were selected. The remaining single-hit probes were selected to ensure an even distribution throughout the genome. To achieve this, the genome sequences were chopped into fragments using EMBOSS (Rice et al., 2000) and one probe was selected from each fragment, excluding the fragments already covered by previously selected probes.

The synthesized probes were used to capture the DNA fragments of the six genotypes. The captured DNA fragments were sequenced using the Illumina HiSeq 3000 platform (100 bp paired-end reads). The probe design, synthesis, library preparation, target enrichment, and sequencing were performed by Rapid Genomics LLC (FL, United States).

Target Capture Efficiency and Coverage of Probes

To evaluate the probe target regions, the sequences of designed probes were aligned to the A + B genomes using Blat following the same criteria as above. The read coverage for probe target regions was assessed. In addition, the relationship between read coverage and target regions' sequence similarities with probes was investigated, which could indicate the influence of alignment identity of probes on capture efficiency. To achieve this, different alignment identity cutoffs were applied to define a hit, including 96, 90, 84, 78, 72, 66, and 60, which correspond to 80, 75, 70, 65, 60, 55, and 50% match of probe sequences to the genome. The coordinates of those hits in the genome were extended 100 bp from both directions (in BED file), which subsequently served as target regions. Bedtools v2.24.0 (intersect) was used for assessing read coverage for target regions. The alignment files for both overall and uniquely mapped reads generated from BWA-mem (Li and Durbin, 2009), as described in section below, were used. Thus, in total seven BED files of target regions under different alignment identity cutoffs, were included for calculating on-target rate and coverage of reads.

RNA-seq and GBS Data Sets

The RNA-seq data of these six genotypes were retrieved from the previous root transcriptome study (Peng et al., 2017a), which were deposited at the Sequence Read Archives (SRA) of the National Center for Biotechnology Information (NCBI, accession number SRP093688, BioProject PRJNA354154, and BioSample SAMN06041692-SAMN06041727). Each genotype had six cDNA libraries, for a total of 36 cDNA libraries for the six samples. In total 403,245,464 read pairs (150 bp) were included for analysis. The raw reads were trimmed with Trimmomatic (Bolger et al., 2014).

The GBS data were obtained for each genotype previously as described by Peng et al. (2017b). The restriction enzyme *Ape*KI was used for removing repetitive regions to reduce genome complexity. A total of 17,408,637 single end reads (100 bp) were obtained (data deposited in the Sequence Read Archives at NCBI under accession number of SRP154150). Raw reads from GBS data were trimmed to 64 bp using Stacks (Catchen et al., 2013).

Read Alignment, SNP Calling and Filtering

The alignment was performed by two general methods (**Table 1**). In the First method, trimmed reads were mapped to A or B genome (A/B) separately, and all mapped reads were used for SNP calling (M1, M4; **Table 1**). In this method, a read coming from the B genome could be erroneously aligned to the A genome, since A and B genomes are quite similar (Bertioli et al., 2016). SNP calling was performed using Samtools 1.3.1 (Li et al., 2009), which was built into the SWEEP pipeline. The homoeologous SNPs generated were further utilized as an anchor for subsequent SNP filtering by using SWEEP and a machine-learning tool SNP-ML (M1). In addition, a haplotype-based genotyping tool HAPLOSWEEP (M4) was also used. So M1 was defined as alignment to A/B genome, using overall aligned reads, and SNP filtering based on SWEEP + SNP-ML and depth. M4 was defined as alignment to A/B genome, using overall aligned reads, and SNP filtering using HAPLOSWEEP (**Table 1**). In the Second method, trimmed reads were mapped to the *in silico* concatenated (A + B) tetraploid genome (concatenated from diploid genomes), and only uniquely mapped reads were used for subsequent analysis (M2, M3, M5; **Table 1**). In this method, only reads having a unique location in the tetraploid genome (according to the aligner) were used. SNP calling was performed using Samtools (Li et al., 2009). SNP filtering was performed by using conventional filtering based on read depth only (M2), SWEEP and SNP-ML (M3), or HAPLOSWEEP (M5). Thus, M2 was defined as alignment to A + B genome, using uniquely mapped reads, and SNP filtering based on depth. M3 was defined as alignment to A + B genome, using uniquely mapped reads, and SNP filtering based on SWEEP + SNP−ML and depth. M5 was defined as alignment to A + B genome, using uniquely mapped reads, and SNP filtering using HAPLOSWEEP (**Table 1**).

When analyzing TES and GBS data, Bowtie2/2.3.4.1 (default – sensitive-local) was used to align reads to A and B genomes separately (for the First method) followed by SNP filtering, which was extensively applied previously in peanut (Clevenger et al., 2017, 2018; Pandey et al., 2017). Due to a low unique mapping rate from Bowtie2, BWA-mem was used for read alignment (for the Second method), which was applied in our previous TES report (Peng et al., 2017b). Uniquely mapped reads from BWA-mem were extracted by filtering off reads with a mapping quality of zero and "XA:Z" tag. When analyzing RNA-seq data, a split aligner Tophat2.1.1 (Kim et al., 2013) was used for both the First method and the Second method, with one mismatch in the 20 bp seed and GFF files supplied (Bertioli et al., 2016). Uniquely mapped reads were extracted by using the tag "NH:i:1" and a mapping quality of "50." "–ultimate" option was used in SWEEP with default settings for other options. For SNP-ML, "-iM peanut_RNA" was used for TES and RNA-seq data, while "-iM peanut_DNA" was used for GBS data. For HAPLOSWEEP, "HAPLOSWEEP_LONGRANGE" was used for TES and RNA-seq data (paired-end reads), and "HAPLOSWEEP" was used for GBS data (single-end reads).

Finally, SNPs called from methods M1, M2, and M3 were filtered based on read depth. A homozygous genotype was called if there were at least four reads supporting either the reference or alternate allele. A heterozygous genotype was called if there were at least two reads supporting the reference and alternate allele, respectively.

Genotyping With the 48K Axiom *Arachis2* Array and Validating SNP Calling Results From NGS Pipelines

The DNA samples of the six parental genotypes were submitted to Affymetrix for genotyping using the recently developed 48K Axiom *Arachis2* array. The genotype calling was performed as previously described (Clevenger et al., 2018). All the SNPs (between PI 262090 and UF 487A) identified from different pipelines used were compared with genotyping results from the SNP array to identify the overlapped or shared SNPs. The polymorphic SNPs (between PI 262090 and UF 487A) identified from those pipelines were considered validated or concordant with the array if they were also polymorphic on the array and had the same genotypes with those called from the NGS methods. The validation or concordance rates for the five SNP analysis pipelines (M1–M5) were subsequently calculated.

RESULTS

Probe Design and Selection for Target Enrichment Sequencing

A total of 199,673 probes were designed for the 3,982 Class I genes and 1,678,459 probes were designed for the 74,753 Class II gene models. After mapping the probe sequences to the genomes (A + B) by Blat, a total of 230,730 probes had a single unique hit (alignment identity ≥96) to the genomes. To avoid any redundancy due to genome sequence duplications, CD-HIT-EST was applied and a total of 219,850 single-hit probes remained. Among the single-hit probes, a total of 20,212 probes corresponding to 2,072 Class I genes, and 9,582 probes covering 907 resistance genes were first selected (**Supplementary Table S1**). In addition, 824 probes with two, three, or four hits to the genomes were also selected since they covered the genes having no single-hit probes. This led to a total of 30,081 probes being selected (**Supplementary Table S1**) covering the Class I and resistance genes.

To select the remaining probes covering Class II gene models, the genome sequences were chopped into 44.3 Kb fragments using EMBOSS and a total of 56,296 fragments were obtained. By excluding the 2,783 fragments that were already covered by the previously selected probes, a total of 24,922 fragments were covered by the remaining single-hit probes. Thus, three fragments were randomly excluded and one probe was selected from each of the remaining 24,919 fragments so that all the selected probes were basically

TABLE 1 | Five different alignment and SNP filtering pipelines.

Method ID	Genome reference	Mapped reads used	SNP filtering		
			SWEEP and SNP-ML	HAPLOSWEEP	Depth-based
M1	A/B	Overall	Yes	No	Yes
M2	A + B	Unique	No	No	Yes
M3	A + B	Unique	Yes	No	Yes
M4	A/B	Overall	No	Yes	No
M5	A + B	Unique	No	Yes	No

"A/B" indicates alignment to A and B genomes separately. "A + B" indicates alignment to a concatenated A and B genome.

evenly distributed throughout the genome. Finally, a total of 55,000 probes (**Supplementary Table S1**) were selected for the TES experiments.

Summary of Sequence Statistics, Trimming and Alignment

On average, there were 14,211,850 paired-end reads (100 bp) per sample obtained from TES, 67,207,577 paired-end reads (150 bp) per sample from RNA-seq, and 2,901,440 single-end reads (100 bp) per sample from GBS (**Supplementary Table S2**). After trimming, 96.89% of the reads remained for TES, 88.29% for RNA-seq, and all reads remained for GBS (reads trimmed to 64 bp). When the trimmed reads were aligned to A/B (A and B genomes separately) genome, on average, the overall mapping rate was more than 96% to either A or B genome for TES, more than 53% for RNA-seq, and more than 82% for GBS. When aligning to the concatenated A + B genome, the average rate of uniquely mapped reads was 51.6% for TES, 50.26% for RNA-seq, and 19.31% for GBS (**Supplementary Table S2**). The low unique mapping rate for GBS was consistent with its short read (64 bp) being used for alignment, in contrast with the 100 bp read length for TES and 150 bp read length for RNA-seq. Certain level of repetitive sequences may exist in the GBS reads, which would also cause low unique mapping rate. Since A and B genomes were quite similar, shorter sequences were less likely to find a unique location when aligned to the A + B genome.

Evaluations of Target Capture Efficiency and Coverage

After mapping probe sequences to the genomes, under the alignment identity cutoff of ≥96, there were 50,580 and 48,275 (91.96 and 87.77% of 55,002) probe target regions covered by reads according to overall and uniquely mapped reads, respectively (**Figure 1A**). By decreasing the alignment identity cutoff, more target regions were available and were covered by reads. Specifically, with an alignment identity between 60 and ∼66, there were still 149,885 and 132,787 (79.57 and 70.49% of 188,369) target regions covered by overall aligned reads and uniquely aligned reads, respectively. The average on-target rates of mapped reads to target regions with an alignment identity ≥96 were 12.82% for overall mapped reads and 16.28% for uniquely mapped reads (**Figure 1B**). The remaining reads were mapped to target regions with a lower alignment identity. If considering all target regions with an alignment identity ≥60, the average on-target rates were 59.81 and 57.69% (**Figure 1B**), respectively. Thus, probes could still capture DNA fragments even with 50% sequence similarity. However, target regions with higher sequence similarities to probes had higher read coverage (**Figure 1C**). Under the alignment identity cutoff of ≥96, the target regions were covered on average 29.86× and 22.05× considering overall and uniquely mapped reads, respectively. It was noteworthy that under cutoff of ≥90, corresponding to ≥75% sequence similarity, the average read coverage was 33.68× and 20.85× for overall and uniquely mapped reads, respectively (**Figure 1C**). The capture efficiency for cutoff 90 was comparable to that of cutoff 96. However, as the alignment identity of the probes was reduced, the average coverage of the reads captured by the probe was reduced as well. Thus, a probe could capture DNA fragments with a high and optimal efficiency if the probe sequence had ≥75% sequence similarity with the fragment sequences.

SNP Calling for NGS Data

The alignment, SNP calling and filtering for three different NGS methods, TES, RNA-seq, and GBS data were performed using five different pipelines (**Table 1**). As there were more polymorphisms between PI 262090 and UF 487A, which were the two original parental lines of E4, E5, E6, and E7, the SNPs identified or validated between these two genotypes were summarized and compared among the five pipelines for the three NGS approaches (**Table 2**). Since these six parental genotypes were not included into the samples for developing the Axiom *Arachis2* array, the randomly overlapped SNPs between the ones identified from the five pipelines and those placed on the array were used for SNP calling cross-validation. For TES data, the largest number of SNPs (22,584) was from M2, followed by M4 (10,157), M1 (7,540), M5 (2,694), and M3 (1,283) (**Table 2**). However, the largest number of homozygous or genome-specific SNPs were identified from M4 (10,157), more than twice the number from M2 (4,438). Similarly, for RNA-seq data, the largest number of SNPs was from M2 (14,684), followed by M1 (1,199), M4 (901), M3 (297), and M5 (288) (**Table 2**). Most homozygous SNPs were also identified from M4 (901), which was higher than M2 (787). For GBS data, 278 SNPs were identified from M4, followed by M2 (171), M1 (161), M5 (15), and M3 (9). Most homozygous SNPs were called from M4 (278) and M2 (37). For

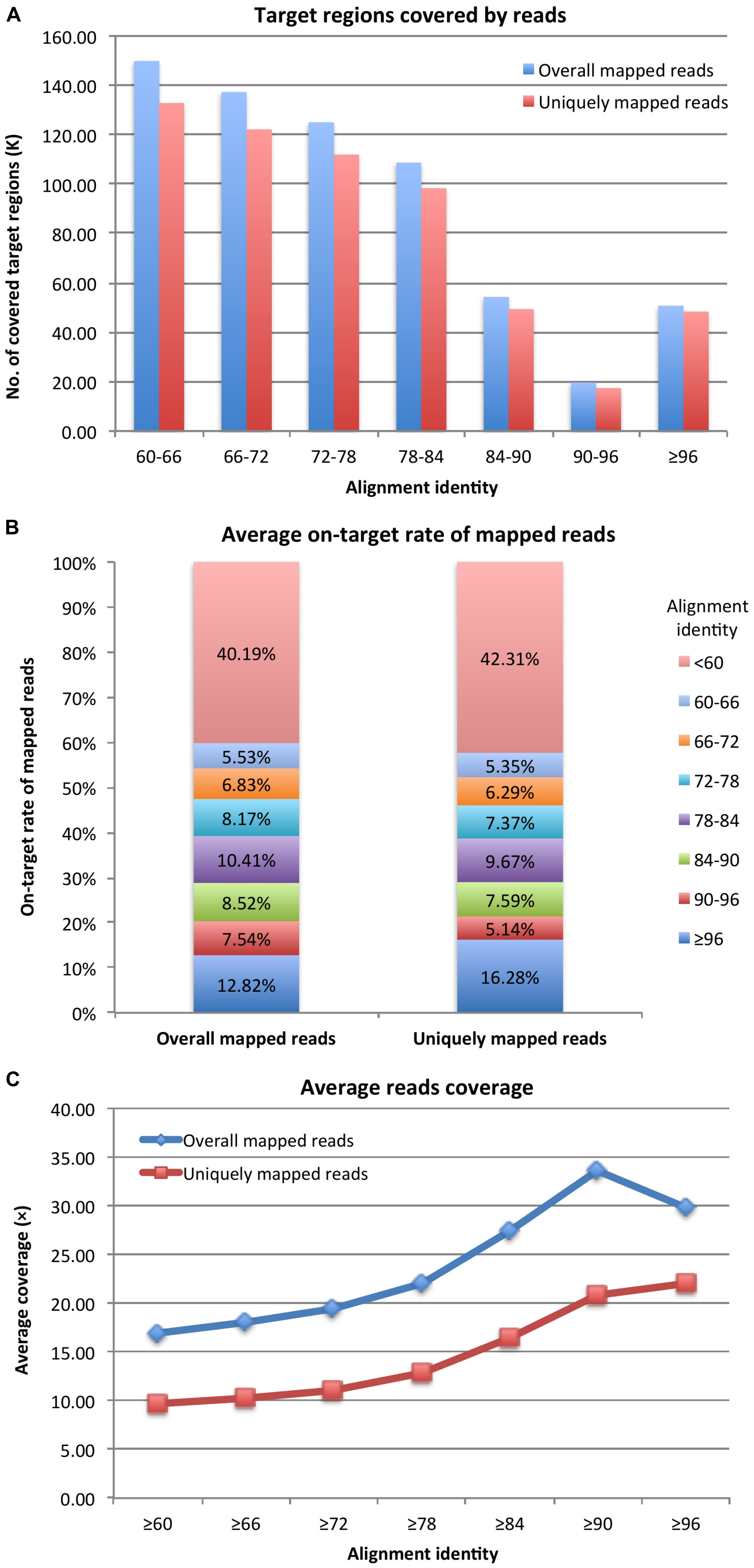


FIGURE 1 | (A) Probe target regions, **(B)** on-target rate of mapped reads, and **(C)** reads coverage for target enrichment sequencing data.

TABLE 2 | Summary of SNPs between PI 262090 and UF 487A from five different methods using target enrichment sequencing RNA sequencing, and genotyping by sequencing data and concordance rate with array-overlapped SNPs.

Data source	SNP analysis method	No. of array-overlapped/total SNPs			No. of concordant SNPs with the array		
		Total	Heterozygous	Homozygous	Total	Heterozygous	Homozygous
TES	M1	88/7,540	86/7,316	2/224	17 (19.32%)	17 (19.77%)	0
	M2	92/22,584	21/18,146	71/4,438	57 (61.96%)	9 (42.86%)	48 (67.61%)
	M3	13/1,283	12/2,938	1/132	5 (38.46%)	4	1
	M4	44/10,157	–	44/10,157	36 (81.82%)	–	36 (81.82%)
	M5	30/2,694	–	30/2,694	23 (76.67%)	–	23 (76.67%)
RNA-seq	M1	30/1,199	26/1,175	4/24	8 (26.67%)	4 (15.38%)	4
	M2	108/14,684	82/13,897	26/787	33 (30.56%)	10 (12.20%)	23 (88.46%)
	M3	13/297	11/285	2/18	3 (23.08%)	2	1
	M4	17/901	–	17/901	14 (82.35%)	–	14 (82.35%)
	M5	9/288	–	9/288	6 (66.67%)	–	6 (66.67%)
GBS	M1	1/161	0/159	1/2	–	–	–
	M2	1/171	0/134	1/37	–	–	–
	M3	0/9	0/9	0/0	–	–	–
	M4	1/278	–	1/278	–	–	–
	M5	0/15	–	0/15	–	–	–

TES indicates target enrichment sequencing; RNA-seq indicates RNA sequencing; GBS indicates genotyping by sequencing.

all three data sources, M4 and M2 identified the largest amount of homozygous SNPs.

Genotyping With the Axiom *Arachis2* Array and the Concordance With NGS Methods

Genotyping using the Axiom *Arachis2* array revealed 23,060 SNP loci with high quality genotypes called for PI 262090 and UF 487A (**Supplementary Table S3**). Of the 23,060 SNP loci, 3,531 SNPs were polymorphic between PI 262090 and UF 487A, including 2,056 homozygous SNPs and 1,475 heterozygous SNPs (**Supplementary Table S3**). After comparison, the SNPs identified using HAPLOSWEEP, either using A/B or A + B as the reference, always had a higher validation rate than other SNP analysis methods based on the aforementioned overlapped SNPs (81.82% for M4, 76.67% for M5) for TES data (**Table 2**). The validation rate was ~79% considering all data points. M2 had a lower concordance rate than M4 and M5, but the concordance rate for homozygous SNPs was 67.61%. All other pipelines either had too few SNPs overlapped with the array or a low concordance rate. Similarly, for the RNA-seq data, M2, M4, and M5 revealed a high concordance rate with the SNP array for homozygous SNPs (**Table 2**). For GBS data, there were too few SNPs from the five pipelines overlapping with those from the SNP array, therefore, they were not included for comparison.

The non-validated SNPs among the overlapped or shared SNP loci were specifically investigated. For M1, most of the non-validated SNPs proved to be polymorphic on the array. However, the genotype calls from sequence data did not match those from the array. Among the 88 overlapped SNPs, 57 (64.77%) of them were called as heterozygous SNPs from sequence data, but as homozygous SNPs from the array. This result showed that M1 was able to identify true polymorphic loci but may not assign a correct genotype due to the alignment of homoeologous reads while the sub-genome specific haplotype cannot be differentiated. In contrast, for the HAPLOSWEEP-based approaches M4 and M5, most of the genotype calls from sequence data matched those from the array (**Table 2**). For the remaining non-validated SNPs from M4 and M5, almost all of them proved to be polymorphic on the array, however, with either PI 262090 or UF 487A showed a heterozygous genotype, which most likely were homoeologous SNPs. Those SNPs on the array could be used as dominant markers. Similarly for M2, the most common non-validated SNP type (22 out of the 92 overlapped SNPs) was classified as a homozygous SNP from sequence data but was called as a heterozygous SNP from the array.

Comparison of Different Platforms

The overall called and cross-validated SNPs among the five pipelines from TES and RNA-seq were further compared (**Figure 2**). For both TES and RNA-seq data, a small proportion (<50%) of called SNPs were shared between M1 and M2, M2 and M4, or between M2 and M5 (**Figures 2A,B**). When comparing the validated SNPs, for TES data, 17 (73.91%; out of 23) of the SNPs from M5 (using A + B as the reference) were already covered by M4 (using A/B as the reference) (**Figure 2C**), both of which applied HAPLOSWEEP. However, only a small proportion (14 out of 57, 24.56%) of the SNPs from M2 overlapped with M4, although both revealed a high validation rate for homozygous SNPs (**Figure 2C**). This was also observed for RNA-seq data, in which only 4 (12.12%) out of 33 SNPs from M2 were covered by M4 (**Figure 2D**). These results showed that M2 and M4/M5 were able to identify different portions of true homozygous SNPs out of the existing true polymorphisms.

The performance of SNP calling and features of the three NGS methods as well as the Axiom *Arachis2* SNP array were compared (**Table 3**). TES revealed the highest amount of homozygous SNPs

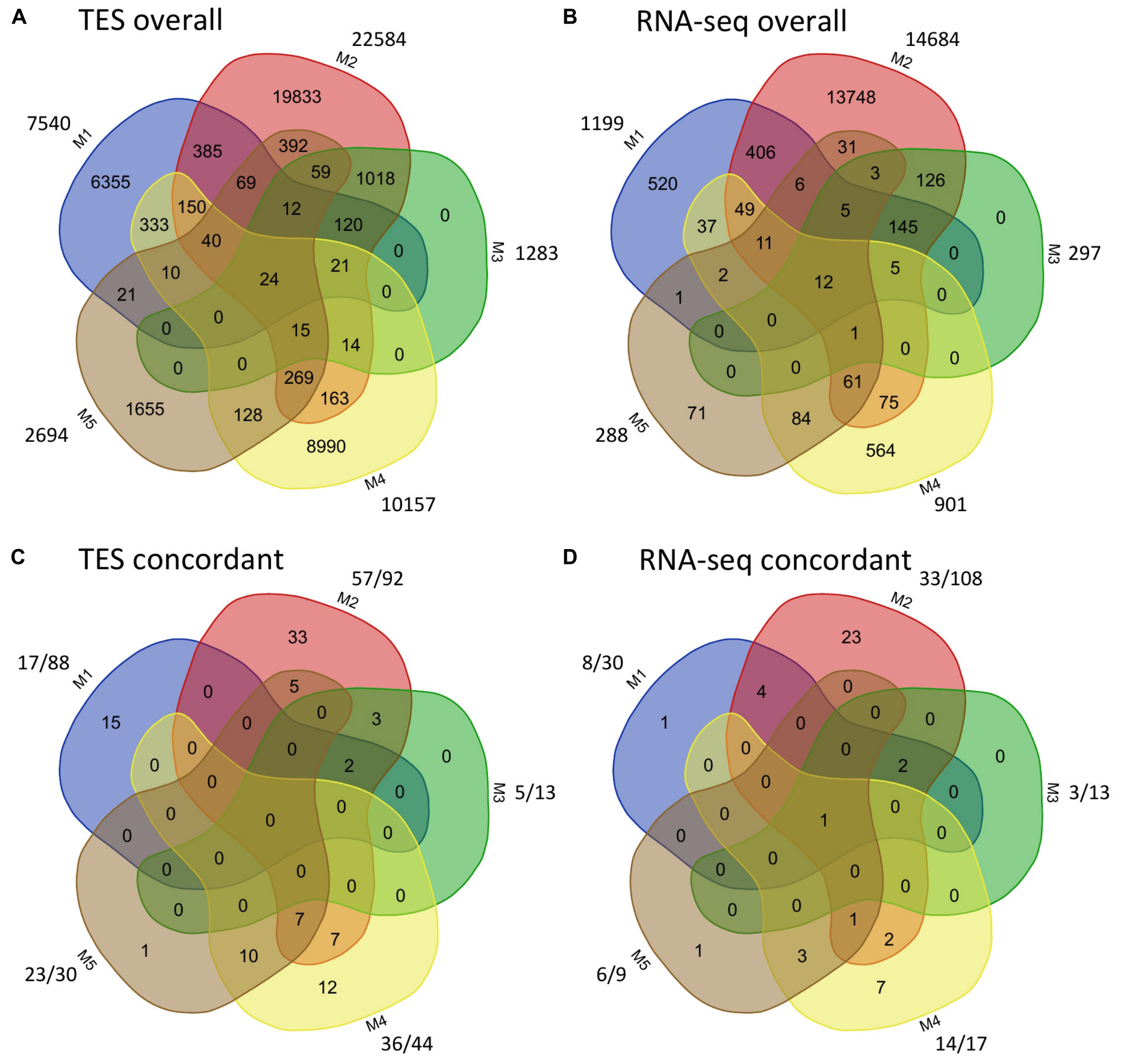


FIGURE 2 | Comparison of identified and concordant SNPs among the five SNP analysis pipelines for target enrichment sequencing and RNA-sequencing data. For panels **(A,B)**, the number outside shows the total number of SNPs identified from each method. For panels **(C,D)**, the number before "/" shows the number of validated SNPs, the number after "/" shows the number of SNPs from each method that are overlapped with the Axiom *Arachis2* SNP array.

(15,947), followed by the Axiom *Arachis2* array (1,887), RNA-seq (1,633), and GBS (312) (**Table 3**). The per sample cost for TES was high compared to other methods, but its per sample per SNP cost was lower than RNA-seq and GBS. However, TES required pre-knowledge of DNA sequences for probe design. The lowest per sample per SNP cost came from the Axiom *Arachis2* array, which also required the least amount of analysis efforts. All three NGS methods required bioinformatics analysis of sequencing data.

Construction of Graphical Maps Containing Polymorphic Regions Between E4 & E5 and E6 & E7

Among the homozygous SNPs between PI 262090 and UF 487A from the Axiom *Arachis2* array, 1,859 (90.68%; out of 2,050 SNPs with high-quality genotypes) were monomorphic between E4 and E5; 1,519 (74.94%; out of 2,027 SNPs with high-quality genotypes) were monomorphic between E6 and E7. By combining the filtered SNPs identified from the three NGS methods as well as those from the Axiom *Arachis2* array, a total of 19,607 non-redundant homozygous SNPs between PI 262090 and UF 487A were obtained. Among those homozygous SNPs, a total of 222 and 1,200 were further obtained between E4 & E5 and E6 & E7, respectively, after filtering. Thus, they were placed on the graphical genotype maps (**Figures 3**, **4**). A total of 75 polymorphic genome regions were obtained for E4 & E5, and 512 polymorphic genome regions were obtained for E6 & E7, which mostly covered and refined those genomic regions revealed by SSR markers (Peng et al., 2018) and potentially harbor genes controlling peanut nodulation. Within the 75 candidate regions of E4 &

TABLE 3 | Comparison of target enrichment sequencing, RNA sequencing, genotyping by sequencing, and the Axiom *Arachis2* array.

Items	TES*	RNA-seq	GBS	Axiom *Arachis2* array
Pre-knowledge of DNA sequences	Yes	No	No	Yes
Efforts of bioinformatics analysis	High	High	High	Low
Price/sample	∼$450	∼$260	∼$35	∼$28
No. of homozygous SNPs identified	15,947	1,633	312	1,887
Per SNP per sample cost	∼$0.0282	∼$0.1592	∼$0.1122	∼$0.0148

*TES indicates target enrichment sequencing; RNA-seq indicates RNA sequencing; GBS indicates genotyping by sequencing. *For 55K probes, including probe design and synthesis.*

E5, there were a total of 67 DEGs and 26 putative orthologous nodulation-related genes, among which *CLE13*, *ENOD16*, *NFR5*, and *NSP2* were also DEGs (**Supplementary Table S4**). Within the 512 candidate regions of E6 & E7, there were a total of 217 DEGs and 39 putative orthologous nodulation-related genes, among which *CLE13*, *ENOD16*, and *RIP1* were also DEGs (**Supplementary Table S4**). Those genes could serve as candidate genes controlling peanut nodulation for further genetic and fine mapping.

DISCUSSION

In this study, we mainly focused on identifying the polymorphic regions between two pairs of sister RILs, E4 & E5, as well as E6 & E7, which are near-isogenic lines. For mapping or fine-mapping the genes controlling nodulation, polymorphic markers differentiating the near-isogenic sister lines are critical and are challenging to develop due to (1) allopolyploid nature of the cultivated peanut and (2) near-isogenic nature of the two pairs of sister lines. Therefore, in this study, we implemented multiple NGS-enabled SNP genotyping methods and SNP calling pipelines to identify reliable and sufficient number of SNP markers.

Single nucleotide polymorphisms have been extensively used for genotyping due to several favorable features such as abundance and high throughput. With the advancement of research in peanut genomics and genetics, especially the advent of reference genomes (Bertioli et al., 2016) and SNP arrays (Clevenger et al., 2017, 2018; Pandey et al., 2017), more choices of SNP genotyping became available for the peanut research community. For genetic mapping studies, WGRS approach can theoretically provide the highest resolution of marker densities. However, for crop species like peanut with a large genome size (∼2.7 Gb), it would still be costly, to have enough sequencing data to meet the requirement of coverage and depth for accurate SNP identification. Alternatively, numerous approaches, such as TES, RNA-seq, and GBS, which reduce the genome complexity by sequencing a partial genome, may be more cost-effective while still able to provide a decent number of markers. In addition, the Axiom *Arachis2* array (Clevenger et al., 2018) is another choice, which involves the least computational analysis efforts. This study utilized six peanut samples to compare SNP identification using sequencing data from different high throughput genotyping methods, TES, RNA-seq, GBS, as well as SNP array. This comparison between the different high throughput genotyping platforms provided an insight into the performance and the number of useful markers that can be generated from each platform. In the past few years, SNP marker development in allotetraploid peanut with highly identical sub-genomes used to be slow due to the presence of homoeologous SNPs (Clevenger et al., 2017). However, with the availability of tools such as SWEEP and HAPLOSWEEP, great progress has been made, which will greatly benefit the whole peanut research community. In addition to these tools, multiple analysis pipelines have also been applied for SNP identification. With so many pipeline options available, a comparison of them was needed to provide a better idea of how they differ from each other and which one outperformed the rest. Current research intended to answer these questions by applying different alignment, SNP calling and filtering methods with different sequencing approaches for SNP identification. Furthermore, the resulting SNPs revealed the polymorphic genomic regions between the sister RILs, which can narrow down the candidate regions harboring genes controlling peanut nodulation, and likely facilitate future genetic mapping and fine mapping of nodulation genes in peanut.

Target Enrichment Sequencing

Unlike RNA-seq and GBS, which focus on genic regions or restriction site-surrounding regions, TES was able to focus on genes or genomic regions of interest. In this approach, the DNA fragments captured by custom-designed probes based on sequence homology were sequenced. Researchers can preferably design probes covering genes of interest. TES was firstly applied in peanut by using probes designed from expressed sequence tags as the sequence source for probe design (Peng et al., 2017b). In the current study, the reference genomes of the two diploid ancestors of cultivated peanut were used for probe design. In order to target symbiosis related and disease resistance related genes in peanut, a total of 20,212 probes were designed to cover all the putative nodulation-related genes and 9,582 probes to cover resistance genes. The remaining ∼24K probes were selected for an even distribution throughout the genome. Therefore, the overall density of the probes was ∼49 Kb/probe given the peanut genome size of 2.7 Gb. Out of the 78,574 peanut gene models, 26,653 (33.9%) of them were tagged by this probe set. This set of TES probes would be useful for not only mapping the genes related to nodulation or disease resistance, but also for genome association analysis of any traits considering the probe density and coverage.

During the probe selection process, single-hit probes were preferably selected, which led to the average unique mapping rate of the five samples to be 51.60%, much higher than our previous report (22.55%; Peng et al., 2017b). In addition, 91.96% of the target regions of current probe set was covered by reads

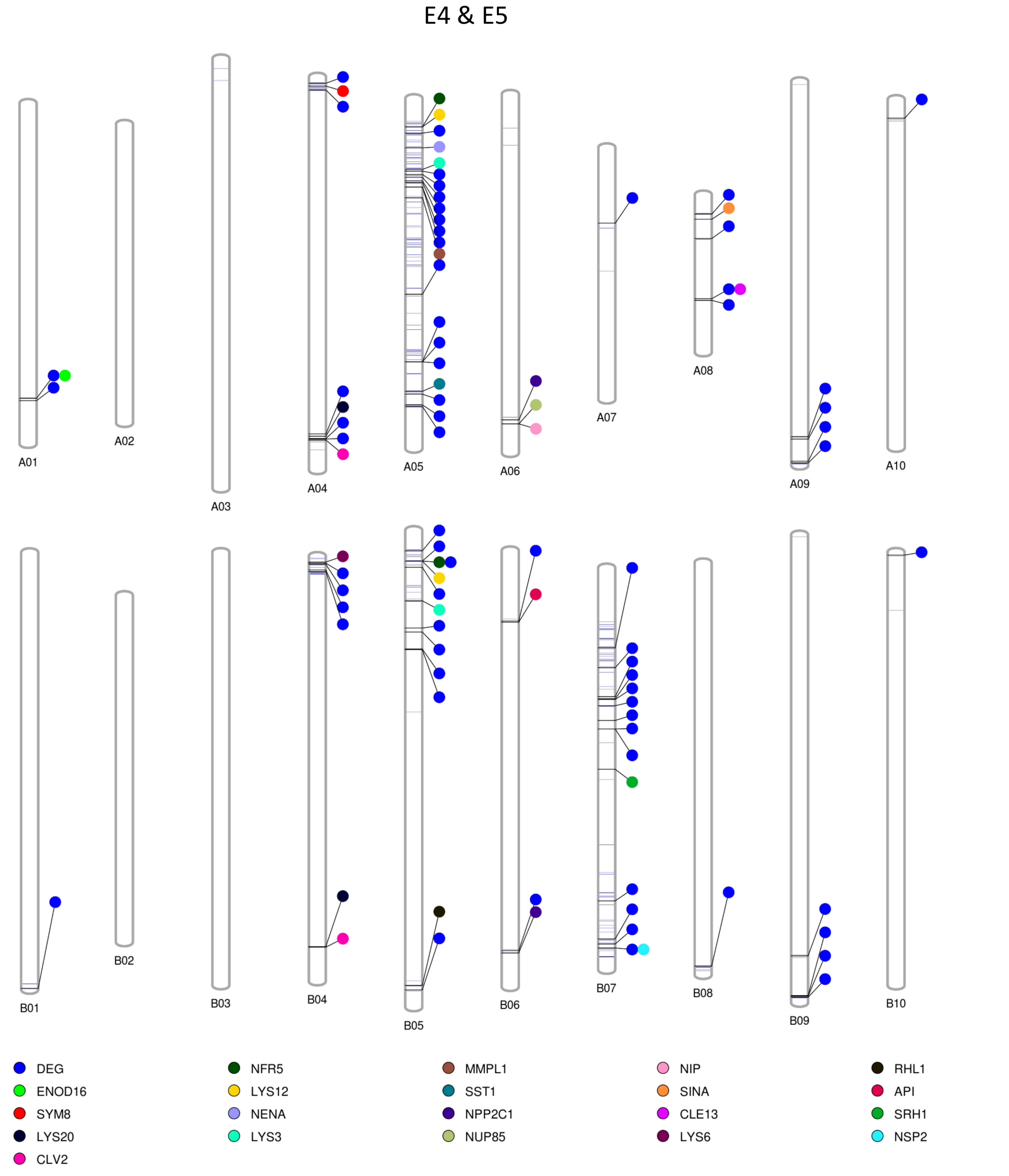


FIGURE 3 | Graphical map showing polymorphic genomic regions between E4 and E5. Each line represents a homozygous SNP. Each circle represents a candidate gene.

with an average depth of 29.86×, which was also much higher than our previous report (average depth <20 × considering 90% of target regions; Peng et al., 2017b). Thus, utilization of the unique hit of probes in the genome is critical to improve the rate of uniquely mapped reads and depth of sequences captured by the probe set. Based on our data, probes can be very efficient in capturing DNA fragments when they have at least 75% sequence similarity with the target fragments (**Figure 1C**). Therefore, when applying TES, we should be aware that off-target capturing would be common specifically for the species with closely related genomes or duplicated regions in the genome.

Comparison of Different NGS Approaches and the Axiom *Arachis2* Array

The three NGS data sources and the Axiom *Arachis2* array identified different numbers of SNPs between PI 262090 and UF 487A. Considering only the homozygous SNPs, TES identified

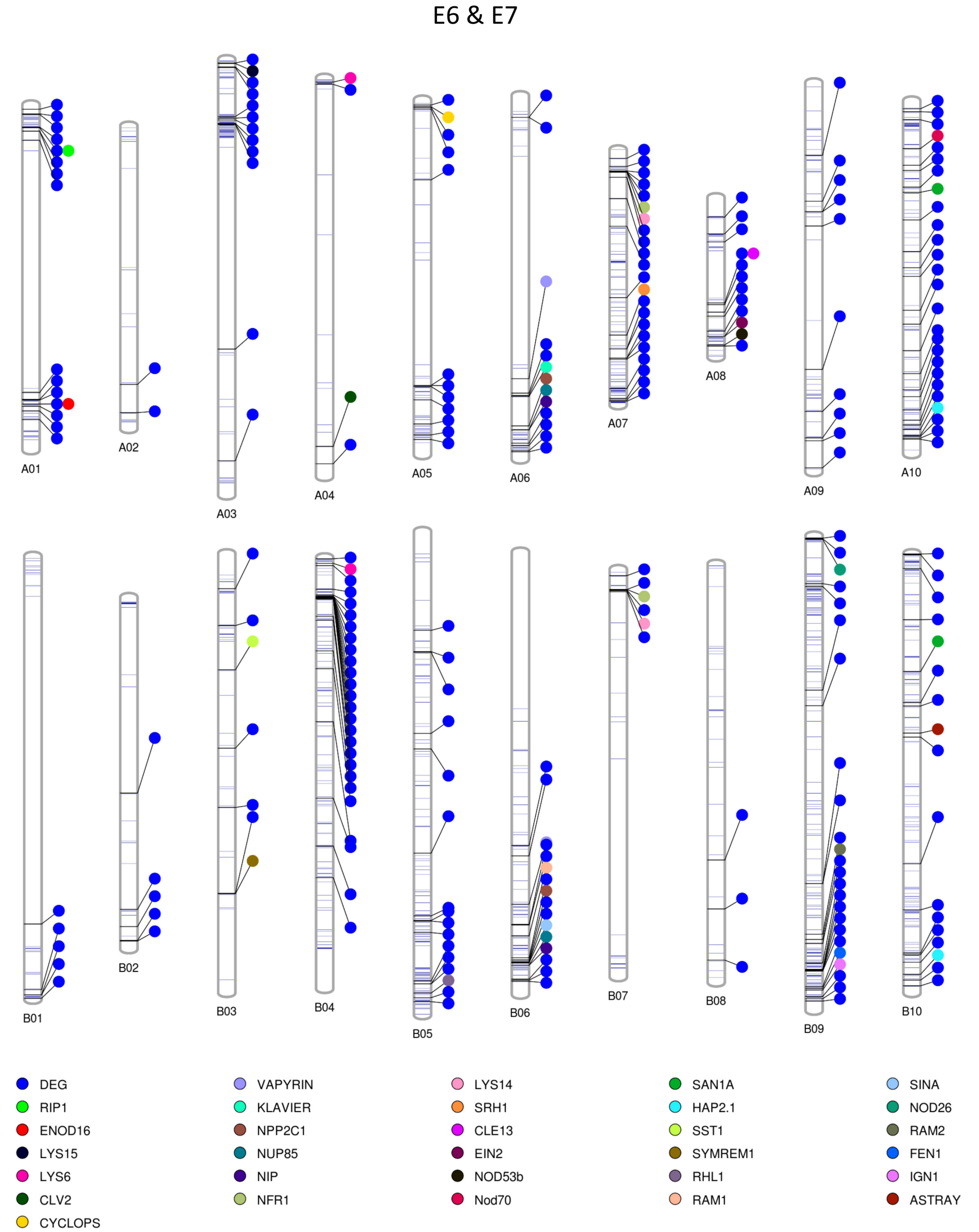


FIGURE 4 | Graphical map showing polymorphic genomic regions between E6 and E7. Each line represents a homozygous SNP. Each circle represents a candidate gene.

the largest number of SNPs, followed by the SNP array, RNA-seq, and GBS (**Table 3**). This could be explained from several perspectives. Firstly, as TES is focused on genomic sequences, more polymorphisms are expected than that from RNA-seq representing the conserved transcribed gene regions. The low number of SNPs from GBS could be explained by the low coverage of sequencing data obtained. As there were only 2,056 homozygous SNPs between PI 262090 and UF 487A obtained from the SNP array, and even fewer SNPs for E4 & E5 and E6 & E7, the Axiom *Arachis2* SNP array may not be suitable for future genotyping of the mapping populations with E4 & E5 and E6 & E7 as the parental lines. TES can be considered as a choice due to the large number of polymorphisms discovered. Moreover, the sample per SNP cost of TES is still low compared to the other NGS methods and comparable to that of the Axiom *Arachis2* SNP array.

Comparison of Different SNP Analysis Pipelines

From the results of comparisons between the five different pipelines for peanut SNP calling, several points can be drawn. (1) The concordance rate of heterozygous SNPs was always low between TES and RNA-seq. This could be caused by false positive SNPs derived from the misalignment of reads from homoeologous regions on the genome. (2) The alignment to A/B genome followed by SWEEP and SNP-ML filtering (M1) revealed a considerably smaller proportion of homozygous SNPs than the alignment to A + B genome followed by traditional filtering (M2), and HAPLOSWEEP approaches M4, and M5. As SWEEP was not able to differentiate haplotypes, by using A/B genome as the reference, a lot of true homozygous SNPs could be called as heterozygous SNPs due to misalignment. (3) M2 revealed a decent concordance rate (67.61%) of homozygous SNPs and could identify new and true polymorphisms that were not found by the HAPLOSWEEP approach. (4) When using HAPLOSWEEP, the alignment to A/B genome (M4) revealed more homozygous SNPs than alignment to the A + B genome (M5), however, M5 could also identify new and true polymorphisms that were not covered by M4. In summary, none of the pipelines above could cover all possible polymorphisms between the genotypes. However, the best option among the five analysis pipelines was to align the reads to A/B genome followed by HAPLOSWEEP, which can yield the highest amount of homozygous SNPs with a high concordance rate with the SNP array, similar to the rate reported in the recent study (74%) (Clevenger et al., 2018). Alternatively, a better choice would be applying multiple pipelines to get non-redundant SNPs. As an example, methods M2 and M4 may complement each other and would yield more homozygous SNPs if both were applied for analysis.

In this study, we used the concatenated A + B genomes from the diploid wild peanut species (Bertioli et al., 2016) as the reference for SNP calling instead of using the tetraploid genomes recently published (Bertioli et al., 2019; Zhuang et al., 2019). One of our main goals in this study was to compare the SNP calling capability using different pipelines and NGS platforms to discover maximum numbers of SNPs in cultivated peanut. This comparison would be reliable as long as the same reference was used for comparison of different platforms or pipelines and the SNPs between the reference and all reads were filtered out. The diploid and tetraploid genomes were highly similar (Bertioli et al., 2019; Zhuang et al., 2019), thus using either genomes as reference would not change the major findings in this study. Particularly, the *Arachis2* SNP array, the tool used for cross-validation of the SNP callings was designed based on the wild diploid genome and the probes designed for TES were also referred to the diploid genomes. Therefore, in this study, a concatenated A + B genome from wild diploid peanut was used for alignment to achieve a good consistency in comparison.

Candidate Genomic Regions Controlling Peanut Nodulation

The two sets of sister RILs used in this study were selected at the F_6 generation, derived from the cross between PI 262090 and UF 487A (Peng et al., 2017a). The Nod+ and Nod− RILs, specifically for E4 and E5, were highly identical. Taking advantage of the nearly isogenic nature between the two pairs of sister RILs with one nodulating and the other non-nodualating, we speculated that polymorphic regions between the sister RILs should harbor any potential candidate genes controlling nodulation. In this study, to identify highly confident homozygous SNPs between the RILs, only the homozygous SNPs polymorphic between PI 262090 and UF 487A as well as between the RILs were included as highly confident SNPs and were placed on the graphical maps. The graphical genotype of these two pairs of RILs allowed us to visualize the polymorphic genome regions harboring candidate genes. The polymorphic regions on the graphical genotype maps could provide guidance for future genetic mapping of nodulation genes in peanut, although these regions were quite big containing a large number of genes since no mapping and fine mapping strategies were applied yet in the current study. We specifically listed out the DEGs involved in nodulation and any orthologs of nodulation related genes as candidates, subsequently obtained a relatively large number of candidates in the genome. These large number of candidate genes was coming from the preliminary comparisons between the two pairs of near-isogenic RILs. Further mapping and fine-mapping strategies should be applied to narrow down and pinpoint the causative genes for non-nodulations in our non-nodulating lines, which will be conducted in a different study.

CONCLUSION

Based on the findings from this study, several suggestions were made for future SNP identification studies in peanut. SNPs included in the Axiom *Arachis2* array were mostly discovered from 21 peanut genotypes, which may not be representative enough to cover all the genome polymorphisms. Axiom *Arachis2* array would be a good choice for genotyping populations developed from or related to the genotypes used for the initial SNP discovery. However, if the populations to be genotyped are not related with the initial genotypes for the development of the Axiom *Arachis2* array, then other NGS approaches should be considered. If genes or genomic regions of interest are to be focused, TES should be preferably considered, since the potential candidate regions can be specifically included for SNP identification. Among the SNP calling pipelines to be used for NGS data analysis, the best performing pipeline is to align the reads to A/B genome followed by SNP filtering using HAPLOSWEEP. To identify a larger number of true homozygous SNPs, other pipelines, such as the alignment to A + B genome with traditional SNP filtering, can be combined with HAPLOSWEEP.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI SRP093688, BioProject PRJNA354154, BioSample SAMN06041692–SAMN06041727, and NCBI SRP154150.

AUTHOR CONTRIBUTIONS

JW conceived the experiments and secured the funding. ZP performed the experiments. ZP and ZZ analyzed the data and drafted the manuscript. JC and DP helped with data analysis. YC and PO-A provided the SNP array data. All authors read and approved the draft.

FUNDING

This work was funded by the Early Career Fund provided by the Institute of Food and Agricultural Sciences (IFAS), University of Florida, National Peanut Board, and Florida Peanut Producers Association. Publication of this article was funded in part by the University of Florida Open Access Publishing Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00222/full#supplementary-material

REFERENCES

Agarwal, G., Clevenger, J., Pandey, M. K., Wang, H., Shasidhar, Y., Chu, Y., et al. (2018). High-density genetic map using whole-genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut. *Plant Biotechnol. J.* 16, 1954–1967. doi: 10.1111/pbi.12930

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., et al. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* 51, 877–884. doi: 10.1038/s41588-019-0405-z

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Chopra, R., Burow, G., Simpson, C. E., Chagoya, J., Mudge, J., and Burow, M. D. (2016). Transcriptome sequencing of diverse peanut (Arachis) wild species and the cultivated species reveals a wealth of untapped genetic variability. *G3* 6, 3825–3836. doi: 10.1534/g3.115.026898

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., and Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Mol. Plant* 8, 831–846. doi: 10.1016/j.molp.2015.02.002

Clevenger, J., Chu, Y., Guimaraes, L. A., Maia, T., Bertioli, D., Leal-Bertioli, S., et al. (2017). Gene expression profiling describes the genetic regulation of *Meloidogyne arenaria* resistance in *Arachis hypogaea* and reveals a candidate gene for resistance. *Sci. Rep.* 7, 1–14.

Clevenger, J. P., Korani, W., Ozias-Akins, P., and Jackson, S. (2018). Haplotype-based genotyping in polyploids. *Front. Plant Sci.* 9:564. doi: 10.3389/fpls.2018.00564

Clevenger, J. P., and Ozias-Akins, P. (2015). SWEEP: a tool for filtering high-quality SNPs in polyploid crops. *G3* 5, 1797–1803. doi: 10.1534/g3.115.019703

Dutta, M., and Reddy, L. (1988). Further studies on genetics of nonnodulation in peanut. *Crop Sci.* 28, 60–62. doi: 10.2135/cropsci1988.0011183x002800010015x

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gallo-Meagher, M., Dashiell, K., and Gorbet, D. (2001). Parental effects in the inheritance of nonnodulation in peanut. *J. Hered.* 92, 86–89. doi: 10.1093/jhered/92.1.86

Gorbet, D., and Burton, J. (1979). A non-nodulating Peanut 1. *Crop Sci.* 19, 727–728. doi: 10.2135/cropsci1979.0011183x001900050045x

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Korani, W., Clevenger, J. P., Chu, Y., and Ozias-Akins, P. (2019). Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *Plant Genome* 12:180023. doi: 10.3835/plantgenome2018.05.0023

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Nigam, S., Dwivedi, S., and Gibbons, R. (1980). "Groundnut breeding at ICRISAT," in *Proceedings of the International Workshop on Groundnut* (Patancheru: ICRISAT Center).

Oldroyd, G. E. (2013). Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. *Nat. Rev. Microbiol.* 11, 252–263. doi: 10.1038/nrmicro2990

Pandey, M. K., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., et al. (2017). Development and evaluation of a high density genotyping 'Axiom_Arachis' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* 7:40577. doi: 10.1038/srep40577

Peng, Z., Fan, W., Wang, L., Paudel, D., Leventini, D., Tillman, B. L., et al. (2017a). Target enrichment sequencing in cultivated peanut (*Arachis hypogaea* L.) using probes designed from transcript sequences. *Mol. Genet. Genomics* 292, 955–965. doi: 10.1007/s00438-017-1327-z

Peng, Z., Liu, F., Wang, L., Zhou, H., Paudel, D., Tan, L., et al. (2017b). Transcriptome profiles reveal gene regulation of peanut (*Arachis hypogaea* L.) nodulation. *Sci. Rep.* 7:40066. doi: 10.1038/srep40066

Peng, Z., Tan, L., López, Y., Maku, J., Liu, F., Zhou, H., et al. (2018). Morphological and genetic characterization of non-nodulating peanut recombinant inbred lines. *Crop Sci.* 58, 540–550. doi: 10.2135/cropsci2017.06.0235

Ren, B., Wang, X., Duan, J., and Ma, J. (2019). Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. *Science* 365, 919–922. doi: 10.1126/science.aav8907

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/s0168-9525(00)02024-2

Rogers, S. O., and Bendich, A. J. (1994). "Extraction of total cellular DNA from plants, algae and fungi," in *Plant Molecular Biology Manual*, eds S. B. Gelvin and R. A. Schilperoort (Dordrecht: Springer), 183–190. doi: 10.1007/978-94-011-0511-8_12

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* 20, 1122–1130. doi: 10.1038/gim.2017.247

Tseng, Y., Tillman, B. L., Peng, Z., and Wang, J. (2016). Identification of major QTLs underlying tomato spotted wilt virus resistance in peanut cultivar Florida-EP TM '113'. *BMC Genet.* 17:128.

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., et al. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics* 15:351. doi: 10.1186/1471-2164-15-351

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* 51, 865–876. doi: 10.1038/s41588-019-0402-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 05 May 2020
doi: 10.3389/fgene.2020.00421

Comparative (Within Species) Genomics of the *Vitis vinifera* L. Terpene Synthase Family to Explore the Impact of Genotypic Variation Using Phased Diploid Genomes

*Samuel Jacobus Smit, Melané Alethea Vivier and Philip Richard Young**

South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Stellenbosch University, Stellenbosch, South Africa

The *Vitis vinifera* L. terpene synthase (*VviTPS*) family was comprehensively annotated on the phased diploid genomes of three closely related cultivars: Cabernet Sauvignon, Carménère and Chardonnay. *VviTPS* gene regions were grouped to chromosomes, with the haplotig assemblies used to identify allelic variants. Functional predictions of the *VviTPS* subfamilies were performed using enzyme active site phylogenies resulting in the putative identification of the initial substrate and cyclization mechanism of VviTPS enzymes. Subsequent groupings into conserved catalytic mechanisms was coupled with an analysis of cultivar-specific gene duplications, resulting in the identification of conserved and unique *VviTPS* clusters. These findings are presented as a collection of interactive networks where any *VviTPS* of interest can be queried through BLAST, allowing for a rapid identification of *VviTPS*-subfamily, enzyme mechanism and degree of connectivity (i.e., extent of duplication). The comparative genomic analyses presented expands our understanding of the *VviTPS* family and provides numerous new gene models from three diploid genomes.

Keywords: terpene, *Vitis vinifera*, functional genomic analysis, gene annotation, carbocation cascade

OPEN ACCESS

Edited by:
Ray Ming,
University of Illinois
at Urbana-Champaign, United States

Reviewed by:
Mario Pezzotti,
University of Verona, Italy
Marianna Fasoli,
E. & J. Gallo Winery, United States

****Correspondence:***
Philip Richard Young
pryoung@sun.ac.za

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *23 December 2019*
Accepted: *03 April 2020*
Published: *05 May 2020*

Citation:
Smit SJ, Vivier MA and Young PR (2020) Comparative (Within Species) Genomics of the Vitis vinifera L. Terpene Synthase Family to Explore the Impact of Genotypic Variation Using Phased Diploid Genomes. Front. Genet. 11:421. doi: 10.3389/fgene.2020.00421

INTRODUCTION

Grapevine has an extensive domestication history that includes various non-*vinifera* hybridizations, resulting in high levels of heterozygosity (Minio et al., 2017). The sequencing of the *Vitis vinifera* cultivar Pinot Noir resulted in the first genome of a woody crop species (Jaillon et al., 2007; Velasco et al., 2007). Inbreeding of Pinot Noir simplified the genome to near homozygosity (93%) which facilitated sequencing of PN40024 (Jaillon et al., 2007). Concurrently a heterozygous clone of Pinot Noir, ENTAV115, was sequenced but difficulties in assembly of the heterozygous and highly repetitive regions resulted in a fragmented genome, limiting its usability (Velasco et al., 2007; Figueroa-Balderas et al., 2019). Continuous improvement over the last decade resulted in numerous assemblies and annotations of the PN40024 reference genome with the latest version (12X.v2 assembly and VCost.v3 annotation) improving the contig coverage and orientation by 14% over the previous assembly (12X.v0) and annotation (v1). However, 2.64 Mbp of contig sequences remain unmapped (chr. 00) while the orientation of numerous mapped contigs remain uncertain (Canaguier et al., 2017).

A combination of crossing (with close relatives as well as non-*vinifera* species) and millennia of propagation have resulted in the expansion of certain *V. vinifera* gene families. Of interest to this study are those linked to volatile organic compounds (VOC) that are often associated with aromatic cultivars. Terpenoids are known to modulate flavor and aroma profiles with monoterpenoids associated with floral and Muscat aromas while a spicy or pepper aroma, in certain wine styles, have been attributed to sesquiterpenoids (Siebert et al., 2008; Skinkis et al., 2008; Wood et al., 2008; Kalua and Boss, 2009; Black et al., 2015; Lin et al., 2019). The genetic potential of a cultivar to form terpenoids is highly variable and modulates the aromatic profile of the derived wine. Wine flavor and aroma is, however, complex and can be influenced by a multitude of factors that not only includes the cultivar but also vinification style, viticultural practices and extent of compound glycosylation (i.e., bound versus free volatiles) (Swiegers et al., 2005; Robinson et al., 2014; Hjelmeland et al., 2015; D'Onofrio et al., 2017). Terpenoids can furthermore be synthesized *de novo* by certain yeasts during fermentation, while other genera are known to liberate bound terpenoids by cleaving the glycosyl bonds (Carrau et al., 2005).

All terpenes consist of the C_5 prenyl diphosphate building blocks isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). These two building blocks arise from the 2-C-methyl-D-erythritol 4-phosphate (MEP) and mevalonate (MVA) pathways that are compartmentalized to the cytosol and plastids, respectively, although metabolic crosstalk between these pathways have been shown (Bloch et al., 1959; Lichtenthaler, 1999; Rohmer, 1999; Bick and Lange, 2003; Hemmerlin et al., 2003). Head-to-tail coupling of IPP and DMAPP results in elongated prenylated substrates that are characteristic to the various known terpene classes. Of particular interest in grapevine due to their volatile flavor and aroma properties, are the C_{10} mono- and C_{15} sesquiterpenes. Monoterpene biosynthesis proceeds through the MEP pathway with geranyl diphosphate (GPP) as the initial substrate with sesquiterpene biosynthesis proceeding through the MVA pathway using farnesyl diphosphate (FPP) and its isomer, nerolidyl diphosphate (NPP) as substrates (Davis and Croteau, 2000). The prenylated substrates can either be ionized or protonated to generate an initial reactive intermediate known as a carbocation, from which a concerted cascade of biochemical reactions proceeds. These reactions include ring-closures, hydride shifts, protonation and deprotonation events and various rearrangements. These cascades, therefore, result in various different carbocation intermediates being formed, subsequent to the initial, allowing for fairly conserved catalytic trajectories that define the enzyme mechanism (Cane, 1990; Davis and Croteau, 2000; Christianson, 2006; Wedler et al., 2015). Sesqui-TPS enzymes are more promiscuous in their product profile due to increased number of orientations that can arise from the added double bond of the FPP substrate, i.e., more possible carbocation intermediates. Enzyme promiscuity is known to be affected by subtle sequence variations in and around the enzyme active site that alter the product specificity or change the enzyme function completely (Li et al., 2013; Drew et al., 2015; Smit et al., 2019). By combining sequence homology of the active site with enzyme functions it is possible to predict how a TPS will interact with its substrate as well as predict the initial step in the carbocation cascade (Durairaj et al., 2019). The more than 40 characterized VviTPS enzymes from different *VviTPS* subfamilies therefore presents an opportunity for grapevine-specific functional predictions using sequence homology and a comprehensive understanding of TPS carbocation mechanisms.

Our current understanding of the grapevine terpene synthase *VviTPS* family is largely based on the PN40024 reference genome. This gene family is extensively duplicated with 152 loci and 69 putatively functional gene models, with the remaining loci being pseudogenes (Jaillon et al., 2007; Martin et al., 2010). However, nearly a third of the family is not mapped to a chromosome (i.e., found mapped to chr. 00), largely due to a lack of contiguity for genomic regions where *VviTPS* genes localize (Canaguier et al., 2017). Furthermore, cultivar-specific gene variations have been shown to impact enzyme function with subtle mutations altering the catalytic mechanism of the enzyme or, most often, rendering a gene non-functional (Drew et al., 2015; Dueholm et al., 2019; Smit et al., 2019).

The reference genome, being near-homozygous, can furthermore not be used to explore potential allelic differences. Allelic differences affecting VviTPS function have, however, been identified using the ENTAV115 heterozygous genome. Although this genome is highly fragmented, it still allowed for the identification of SNPs in *VviTPS24* that alters the catalytic mechanism from producing selinene-type sesquiterpenes to α-guaiene, the key precursors for synthesis of the rotundone sesquiterpene (associated with pepper aromas in wine) (Drew et al., 2015). Cultivar-specific *VviTPS* functions have been shown in a limited number of cultivars (Martin et al., 2010; Drew et al., 2015; Dueholm et al., 2019; Smit et al., 2019). Extrapolating this to the more than 6000 grapevine accessions planted worldwide (This et al., 2006) suggests extensive *VviTPS* diversity, with the PN40024 genome sequence likely representing only a fraction of the genetic potential.

The recently available draft diploid genome assemblies for grapevine provide extensive new genomic information that can be utilized to explore cultivar-specific *VviTPS* variation to understand structure-function relationships (i.e., gene-protein-terpene) for terpene biosynthesis. In this study Cabernet Sauvignon (CS), Carménère (CR), and Chardonnay (CH) were selected for this purpose as they were sequenced and assembled using the same technology: Pacific Biosciences Single Molecule Real Time Sequencing (PacBio-SMRT) sequencing with FALCON-UNZIP phased assembly (Chin et al., 2016; Roach et al., 2018; Minio et al., 2019). The PacBio-SMRT platform allows for long-read sequencing (> 30 kb), resulting in highly contiguous reads that are easier to assemble, but with a greater error rate (7–15%) than short-read sequencing (Rhoads and Au, 2015). The latter limitation is, however, overcome by the greater read-depth (> 115X versus 12X for the reference genome) (Chin et al., 2016; Figueroa-Balderas et al., 2019). Phased assembly with FALCON-UNZIP allowed for haplotype resolution, resulting in affectively two assemblies: the primary assembly, consisting of highly contiguous pseudo-molecules that contain both haplotypes, and the haplotig assembly, consisting of shorter phased reads that represent alternate alleles (Chin et al., 2016; Minio et al., 2017). The differences in assembly approach between the latest diploid grapevine genome and the PN40024 reference genome is illustrated in **Figure 1**. The diploid genomes of CS, CR and CH are highly contiguous and more complete than the PN40024 reference genome (N50 of 0.94–2.17 Mbp versus 0.103 Mbp). The phased diploid genomes therefore allow for genomic sequence data that captures homo- and heterozygous gene regions as well is hemizygous regions (gene regions unique to a haplotype) (Jaillon et al., 2007; Chin et al., 2016; Roach et al., 2018; Minio et al., 2019). The diploid genomes sizes are, however, inflated (more than double the haploid genomes) due to haplotype regions being missed in regions of high heterozygosity, resulting in the haplotypes being incorrectly assigned to the primary assembly (Minio et al., 2017; Figueroa-Balderas et al., 2019).

Focusing on the *VviTPS* family, the aim was to evaluate and correct gene models, where necessary, and then explore the extent of haplotype and genotype variations using the phased diploid draft assemblies. An in-depth analysis of the three genomes ultimately resulted in a significant extension of current knowledge on the *VviTPS* family; which includes chromosome groupings, functional prediction (which includes TPS-subfamily, initial substrate and cyclization mechanisms), cultivar-specific duplication analysis and identification of conserved VviTPS functions. Interactive networks were constructed for gene duplication and genotype/haplotype variations, making the data easily accessible. These networks can be queried using BLAST and all relevant VviTPS information interactively accessed in the respective networks.

METHODOLOGY

Genome Assemblies and Annotations Utilized

Genomes for *V. vinifera* cultivars (Jaillon et al., 2007; Chin et al., 2016; Minio et al., 2017, 2019; Roach et al., 2018) listed in **Table 1** where downloaded from the listed repositories. PN40024 12X.v2 assembly and VCost.v3 (V3) annotation was used as the reference genome (Jaillon et al., 2007; Canaguier et al., 2017). The GFF3 annotation for the terpene synthase family[1] was used for *VviTPS* positioning on the reference genome. PN40024 *VviTPS* sequences identified and curated by Martin et al. (2010) were retrieved from FLAGdb + + (Dèrozier et al., 2011).

The domestication history of these cultivars was inferred by using the *Vitis* International Variety Catalog[2] and domestication histories described by Myles et al. (2011) and Minio et al. (2019).

Identification and Annotation of *VviTPS* Gene Regions on the Diploid Genomes

The Exonerate tool (Slater and Birney, 2005) was used to identify VviTPS-like regions on the primary and haplotig assemblies of the respective diploid genomes (**Table 1**). PN *VviTPS* gene models served as query sequences with the exonerate parameters set to the est2genome model, percentage of the maximal score set at 90% and intron size limited to 3000 bp. The est2genome model parameter employs a gapped alignment algorithm of all *VviTPS* reference sequences to query all primary contig and haplotig sequences for the presence of a *VviTPS*-like gene region. A detailed explanation of the Exonerate analysis can be found in **Supplementary Data Sheet 1**. Exonerate computations were performed using the Stellenbosch University Central Analytical Facilities' HPC2: http://www.sun.ac.za/hpc. Exonerate-gff outputs were annotated on the respective genome contigs and manually curated with CLC Main Workbench 7 (CLC Bio-Qiagen, Denmark) to identify hit regions with the greatest coverage and highest mapping score.

The identified gene regions were compared with the computational annotations reported for the respective genomes in **Table 1**. Each identified gene region was assigned a unique accession consisting of a two-letter cultivar code (**Table 1**) followed by a sequential TPS number. When automated annotations for the respective genomes (**Table 1**) where congruent with the annotation generated in this study, the annotation specific locus ID was maintained as the parent ID in the annotation file. Annotated coding sequences of these congruent regions were maintained as far as possible, but manual correction of several gene regions was necessary with such corrections noted in the annotation file. Gene regions lacking a parent ID indicates a newly annotated region. Partial genes were noted when there were four or less exons and a lack of start and/or stop codons. Genes were considered to be complete if a start and stop codon was present at the terminal ends and the exon number was greater than four. Complete genes were subsequently evaluated for the presence of a reading frame, with genes lacking a full-length open reading frame (fl-ORF) tagged as disrupted (d-ORF), with the disruption being either a premature stop, or a frameshift (insertion-deletion) mutation.

Protein sequences were derived for the fl-ORF's and the terpene synthase N-terminal (PF01397) and C-terminal (PF03936) domains predicted using the PFam Domain Search function of CLC Main Workbench 7 (CLC Bio-Qiagen, Denmark). The motif search function CLC Main Workbench 7 was used to identify motifs characteristic to TPS proteins (Starks, 1997; Williams et al., 1998; Rynkiewicz et al., 2002; Gao et al., 2012; Durairaj et al., 2019).

Putative Identification of Duplicated Gene Regions

A BLASTn alignment (Altschul et al., 1997; Camacho et al., 2009) of complete gene regions for each cultivar was performed and duplications identified by calculating the identity (I') using the formula described by Li et al. (2001), with I being the number of identities and gaps, n the aligned length and L the total length of the query and subject sequences, respectively. For BLASTn analyses of primary-to-primary and haplotig-to-haplotig complete genes, an E-value of 1e-5 was used, with the maximum number of alignments (max-hsps) limited to 5 and number of aligned sequences (max-target-seqs) set to 10. The latter

[1] https://urgi.versailles.inra.fr/Species/Vitis/Annotations

[2] www.vivc.de (accessed November 18, 2019)

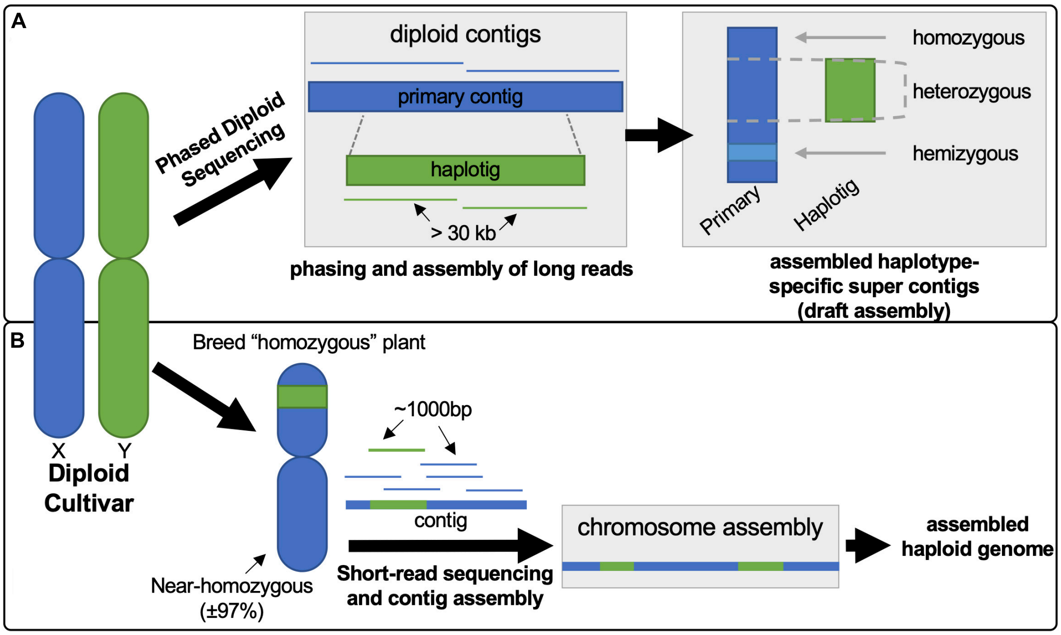


FIGURE 1 | Illustration contrasting the differences in approach between phased diploid **(A)** and short-read sequencing **(B)** to generate the diploid and PN40024 reference genomes, respectively. **(A)** PacBio long-reads (> 30 kb) undergo phasing during the assembly, using FALCON-UNZIP, resulting in the generation of a pseudomolecule known as a primary contig. This primary contig region represents both possible haplotypes. The phasing algorithm identifies reads with high heterozygosity to call an alternate genomic region (haplotig) which is assembled separately allowing for the generation of haplotype regions. **(B)** The reference genome was generated from an inbred clone of Pinot Noir (PN40024), reducing genomic complexity to near homozygosity. Assembly of the short-reads, generation of contigs and subsequent mapping/assembly to chromosomes lead to the PN40024 reference genome.

two parameters were set to 1 when haplotigs-to-primary alignments were performed.

Rapid Assembly of Contigs

Chromosome positions of *VviTPS*-like gene regions were inferred by mapping all *VviTPS* containing contigs to the PN reference genome using rapid reference-guided assembly (RaGOO) (Canaguier et al., 2017; Alonge et al., 2019). The RaGOO parameters for chimera breaking (-b), structural variant calling (-s) and a gap padding (-g) of 200 were used with unplaced contigs not assembled to a random chromosome. The random pseudo-molecule (chr. 00) of the reference genome was not included for RaGOO assemblies. The output of this cultivar-specific all-against-all assembly was used to group contigs according to their highest scoring PN40024 chromosome, followed by chromosome specific contig assembly. The respective RaGOO outputs were visualized using the contig alignment function of the Alvis tool (Martin and Leggett, 2019).

Functional Annotation of *VviTPS* Genes

Multiple sequence alignments (MSA) and phylogenetic tree constructions were performed in the CLC Main Workbench 7 (CLC Bio-Qiagen, Denmark). For nucleotide alignments the ClustalO algorithm was used while the MUSCLE algorithm was used for protein sequences. Phylogenetic trees were constructed with UPGMA, Jukes-Cantor as distance measure and 100 bootstrap replicates (Jukes and Cantor, 1969; Edgar, 2004a,b). MSA's were performed at the nucleotide level using the 152 *VviTPS* gDNA and mRNA sequences predicted by Martin et al. (2010) as reference. Phylogenetic position relative to PN40024 gDNA sequences were used to group gene regions into TPS-subfamilies (Bohlmann et al., 1998; Martin et al., 2010) with the eulerr R package (R Core Team, 2013; Larsson, 2019) used to visualize the data.

Protein sequence phylogenies with characterized grapevine TPSs (**Supplementary Table 1**) were used to group proteins into TPS-subfamilies. For the TPS-a subfamily, the active site region was identified as described by Durairaj et al. (2019)

TABLE 1 | Genome assemblies and annotations utilized.

Genome	Assembly type	Annotation version	Repository
PN40024 12X.v2 (PN)	Haploid	VCost.v3	https://urgi.versailles.inra.fr/Species/Vitis
Cabernet sauvignon (CS)	Diploid	V1	http://cantulab.github.io/data.html
Carménère (CR)	Diploid	V1	http://cantulab.github.io/data.html
Chardonnay (CH)	Diploid	V1	https://doi.org/10.5281/zenodo.1480037

and aligned as described earlier. This active site phylogeny and the Database of Characterized Plant Sesquiterpene Synthases (Durairaj et al., 2019) was used to divide TPS-a members into groups based on their parent cation and first cyclization. For the TPS-b subfamily a similar approach to Durairaj et al. (2019) was applied where only the active site region between the C-terminal metal binding motifs, if present, were aligned. The product profiles of TPS-b members (Martin et al., 2010) were used to predict a mono-TPS reaction mechanism (Williams et al., 1998; Davis and Croteau, 2000; Schwab et al., 2001; Hyatt et al., 2007; Schwab and Wüst, 2015; Xu et al., 2018) and categorize proteins according to their initial carbocation intermediate (terpinyl or linalyl cation). The latter was further subcategorised by considering whether or not quenching occurs before deprotonation. The TPS-g subfamily was subcategorised using the full-length protein alignment and phylogenetic position relative to functional proteins (Martin et al., 2010).

Finding Homologous Proteins Between Cultivars

The cluster function of MMseqs2 (Steinegger and Söding, 2017) was used for all-against-all clustering of proteins with the following parameters: bidirectional alignment coverage mode with a minimum coverage of 85%, minimum sequence identity of 75%, *E*-value of 1e-5 and greedy clustering (cluster-mode 2). Representative sequences from the clustering were extracted as described in **Supplementary Data Sheet 2**.

Network Construction

Cytoscape v3.7.2 (Shannon et al., 2003) was used to construct all networks presented in this study with the data generated from the aforementioned methodologies used for node and edge metadata.

RESULTS

Relatedness of the Genomes

The domestication history of grapevine (Myles et al., 2011) and available pedigree information (Maul and Töpfer, 2015) shows that CR and CS have a common parent while Pinot Noir is a parent to CH. All cultivars share Traminer as an ancestor. The relatedness (pedigree) of cultivars used for genomes discussed in this study is shown in **Supplementary Figure 1**.

Diploid Genome VviTPS-Like Gene Regions

Nearly all of the diploid contigs annotated with a *VviTPS* could be assigned to a reference chromosome using RaGOO, with the exception of 1 CS and 2 CR contigs. The position of the mapped contigs were congruent to *VviTPS* containing chromosomes of the reference genome (Martin et al., 2010). The RaGOO grouping scores per chromosome (**Supplementary Table 2**) ranged between 53% and 97% with an average of 75%, indicating that the contigs could be placed on a chromosome with an acceptable level of confidence. However, the exact position on a chromosome could not be accurately estimated, as evident by the location scores, reflecting a low level of collinearity to the reference genome (**Supplementary Table 2**). Contig alignments to the reference genome using Alvis (**Supplementary Data Sheet 3**) clearly illustrates the extent of discontiguity when mapping the phased diploid contigs to the reference genome.

The Euler graphs in **Figure 2** show *VviTPS* subfamily members per chromosome for the diploid assemblies with PN40024 as a reference. Despite the latest assembly improvements for PN40024, a large number *VviTPS* genes are yet to be assembled to a chromosome, reflected by the "unplaced" genes in **Figure 2**. The diploid assemblies showed an inverse proportional relationship between unplaced and chr. 10 genes relative to PN40024, indicating that long read sequencing has overcome, to a large extent, the unresolved location of chr. 10 *VviTPS* genes. 28 *VviTPS* genes for CH and 41 for CR and CS, respectively, were placed on chr. 10, compared to a single gene on PN40024 (Martin et al., 2010; Canaguier et al., 2017). The majority of these genes are homologous to members of the PN40024 TPS-g subfamily, as illustrated by the gDNA phylogeny in **Supplementary Figure 2**. Furthermore, CR had more than three times the number of genes on chr. 01, 07 and −08 than CS, CH or PN. In agreement with the reference genome, the majority of TPS-a genes are located on chr. 18 and −19 with nearly all TPS-b genes on chr. 13.

The distribution of complete and partial gene regions on the primary and haplotig assemblies is shown in **Figure 3A**. Complete gene regions were sub-categorized into fl-ORF or d-ORF, with the latter representing regions that can also be considered as pseudogenes. Although CR had the greatest number of *VviTPS*-like regions (243), only 49% of these regions encode for a putative fl-ORF, shown in **Figure 3B**, with 84% of the complete genes being duplicated (**Figure 3C**). CS and CH had a similar number of *VviTPS*-like regions (203 and 192, respectively), with CH showing the greatest proportion of fl-ORF (77%) of all three cultivars (**Figure 3B**). CS and CH *VviTPS* families are also extensively duplicated, however, ∼30% of their complete *VviTPS* genes were hemizygous (**Figure 3C**).

Despite the diploid genomes being unassembled, the size and contiguity of the phased diploid contig assemblies allowed for the extent of gene duplications to be investigated, as illustrated by the cultivar specific networks in **Figures 4A–C**. Gene regions with an identity score (I') greater than 80% were considered to be duplicated with those localizing to the same contig considered to be tandem duplicates. Tentative duplications show genes that are not on the same contig (i.e., possible genome wide duplications). **Supplementary Figure 3** shows an alternative node coloring for the aforementioned figure, illustrating their chromosomal localization. The duplication distribution in the edge interactions graph (**Figure 4D**) gives an estimation of the homozygosity for each cultivar, with CS showing the greatest percentage (32%) of haplotype edge connections, i.e., potential allelic variants.

Functional Annotation of the *VviTPS-a* and *-b* Subfamilies

Protein sequences derived from fl-ORFs and subsequent phylogenetic similarity to known functional VviTPS enzymes

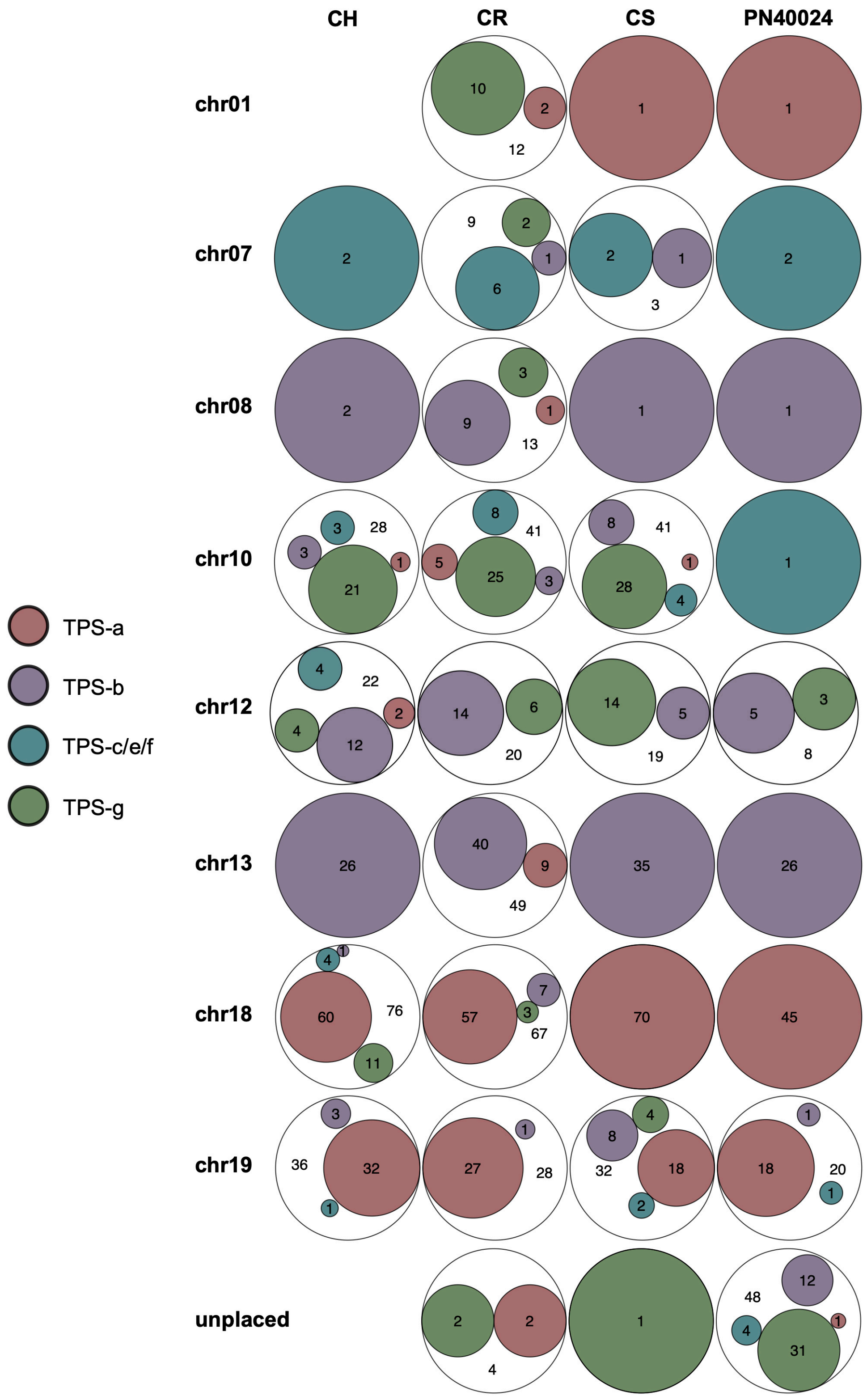


FIGURE 2 | Euler diagrams summarizing the chromosome specific distribution of *VviTPS*-subfamilies for each of the diploid genomes: Cabernet Sauvignon (CS), Carménère (CR), and Chardonnay (CH) as well as the Martin et al. (2010) annotation of PN40024 (PN). The legend shows the different *VviTPS* subfamilies, proportionally sized within the Euler diagrams to reflect the total number of *VviTPS*-like gene regions per cultivar and chromosome.

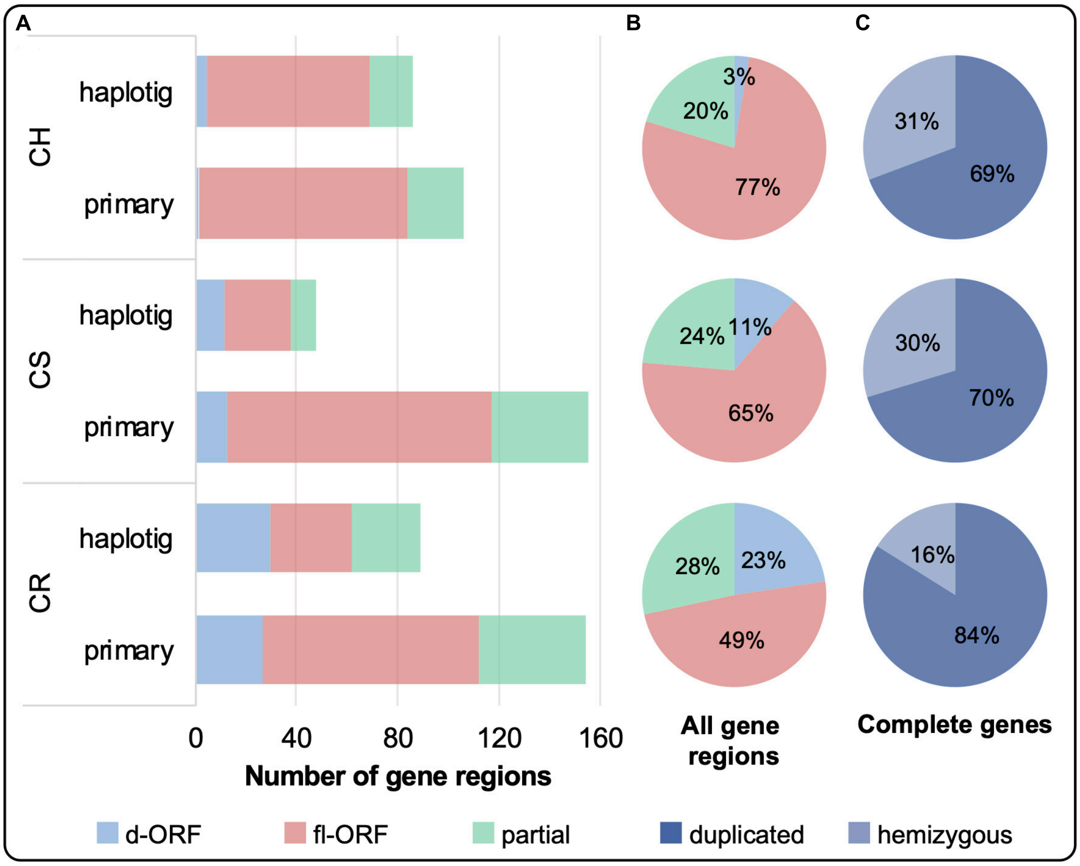


FIGURE 3 | (A) The total number of *VviTPS*-like gene regions on the primary contigs and haplotigs is shown for the draft diploid genomes of Cabernet Sauvignon (CS), Carménère (CR), and Chardonnay (CH). **(A)** The total number of *VviTPS*-like gene regions were further classified by the type of open reading frame (ORF): disrupted (d-ORF) contain frameshifts and/or premature stop codons that render the gene non-functional; full-length (fl-ORF) are predicted to be functional; and partial genes that have four less exons (i.e., pseudogenes). The combined percentage distribution of these ORFs across the haplotypes is shown in **(B)**. The percentage of complete gene regions, the sum of fl-ORF and d-ORF, that are duplicated (degree of similarity (*I'*) > 80%) or hemizygous is shown in **(C)**.

clearly separate the proteins into subfamilies, illustrated in **Supplementary Figure 4**. The VviTPS-a, -b and -g subfamilies represent the majority of putative proteins and were subsequently analyzed in a family specific manner to predict their function.

The VviTPS-a subfamily separates into three major groups based on the initial substrate (FPP and/or NPP) utilized, illustrated in **Figures 5A,B**. Two acyclic subgroups were associated with each of these substrates. With the exception of the acyclic sesquiterpenes, all enzymes that use NPP as sole substrate will proceed through an initial 1,6-cyclization of the nerolidyl cation (Davis and Croteau, 2000). Reactions mechanisms that proceed from FPP formed three distinct clades, indicated by the red triangles, with each clade showing a group for 1,10- and 1,11-cyclizations. Acyclic sesquiterpenes and those that require 1,11-cyclization showed commonality in clade 1 that is distinct from the 1,10-cyclization group. Clade 2 showed three distinct groups with a unique subgroup consisting of both 1,10 and 1,11-cyclization enzymes. The third clade had a number of enzymes that could not be definitively placed into a cyclization group but, as with clade 2, showed clear separation between the 1,10 and 1,11 cyclization mechanisms. The putatively functional *VviTPS-a* genes for each cultivar ranged between 41 and 74, as illustrated by the bar graph in **Figure 5**. Furthermore, the number of genes associated with the respective carbocation cascades (**Figure 5B**) differs between cultivars. The 1,10 and 1,11 cyclization of FPP represents the majority of reaction mechanisms in all cultivars. Enzymes predicted to form acyclic sesquiterpenes were limited to between 2 and 4, while 1,6 cyclization of NPP represents less than a third of the predicted mechanisms.

Although the VviTPS-b subfamily utilizes a single substrate for monoterpene biosynthesis, enzymes could still be grouped into distinct reaction mechanisms, illustrated in **Figures 6A,B** where the cyclic reaction mechanism is referred to as TPS-b Type I while the acyclic mechanism is referred to as TPS-b Type II. Type II enzymes however, formed three distinct clades, of which two are for the single product enzymes associated with linalool (red branch) and ocimene biosynthesis (blue branch), VvPNRLin and VvGwBOci/VvCSbOci, respectively (Martin et al.,

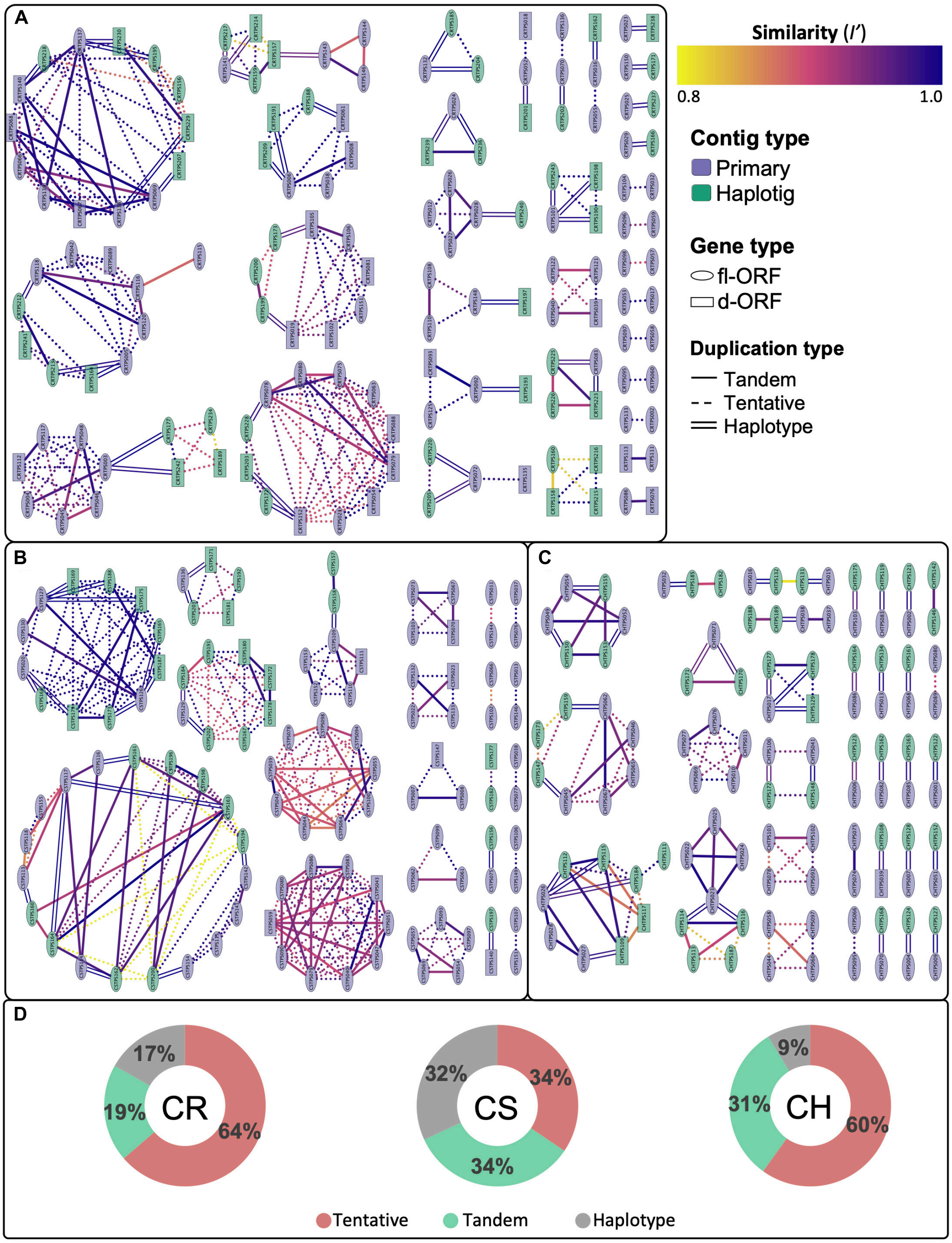


FIGURE 4 | Cytoscape network illustrating the connectedness and degree of similarity (*I*′) for duplicated complete genes for **(A)** Carménère (CR), **(B)** Cabernet Sauvignon (CS) and **(C)** Chardonnay (CH). Nodes represent *VviTPS* genes and are connected by edges, signifying homology of *I*′ > 80%. Complete genes are grouped into those with a full-length or disrupted open reading frames (fl-ORF or d-ORF) for the cultivar-specific haplotypes. The type of edge interactions are further categorized as tandem duplicates if the gene is present on the same contig; haplotype duplications are on primary contigs and haplotigs that localize to the same chromosome and were inferred from RaGOO assemblies to PN40024; with tentative duplications showing genes with high homology that cannot be defined by the two previous groupings. The total percentage contribution of these groupings is shown in **(D)**.

2010). The third Type II clade (light green branch) is represented by a single functional enzyme (VvCSbOciM) that produces 98% acyclic monoterpenes, (*E*)-beta-ocimene and myrcene, and minor amount of the cyclic monoterpene pinene (Martin et al., 2010). This clade is also the largest in all three cultivars, as shown by the bar graph in **Figure 6**, and is closely related to a clade of multiproduct Type I mono-TPS enzymes (yellow branch). The phylogenetic distribution and predicted reaction mechanisms therefore show that the majority of mono-TPS genes will produce both cyclic and acyclic monoterpenes.

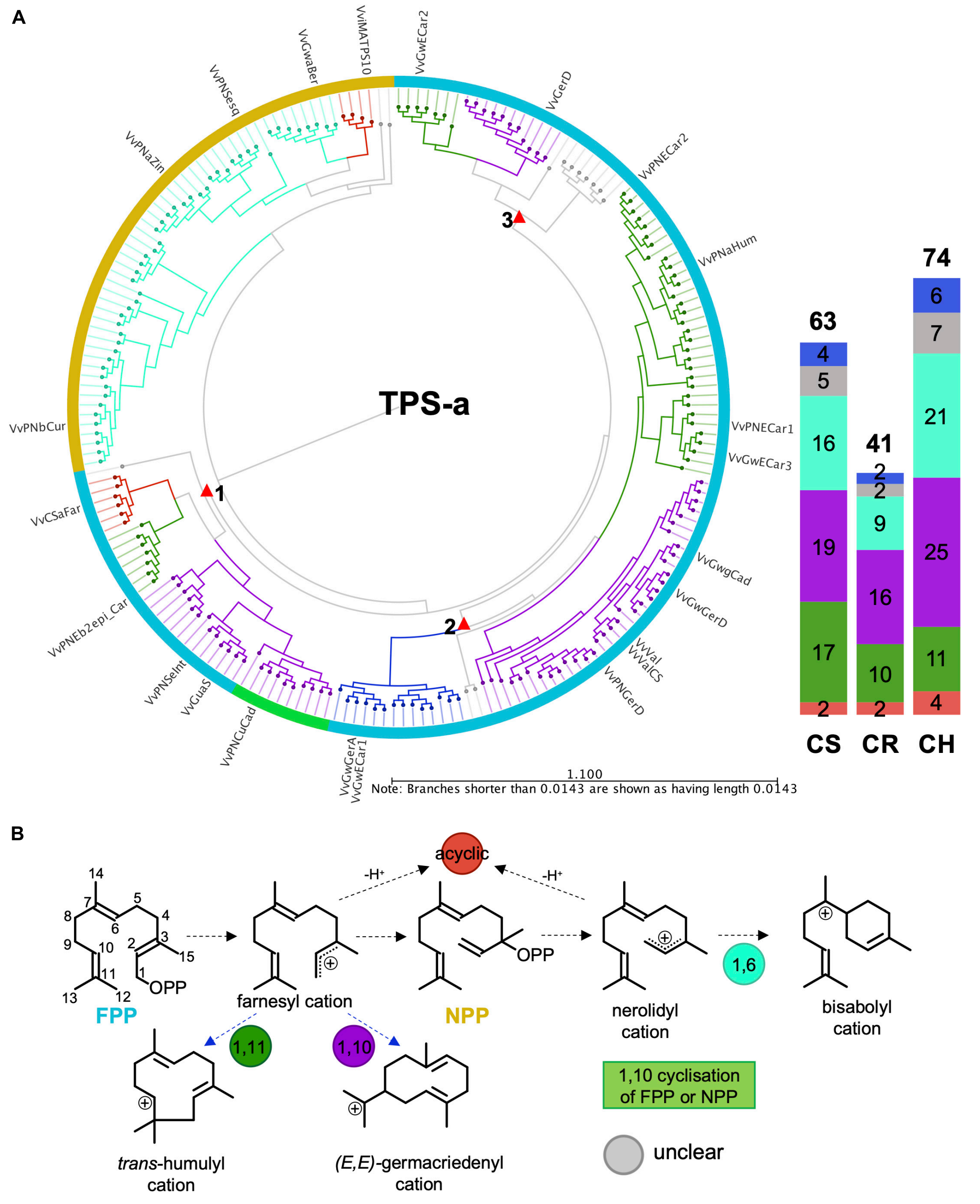


FIGURE 5 | (A) Active site phylogeny of putative VviTPS-a proteins from the diploid genomes and functionally characterized enzymes (outer labels) were used to predict the initial substrate (outer colored ring) and cyclization mechanism (branch color). The red triangles indicate subclades for enzymes that utilize farnesyl diphosphate (FPP) as initial substrate, shown by the blue outer ring. The yellow outer ring represents enzymes that utilize nerolidyl diphosphate (NPP) as initial substrate. Both FPP and NPP can be ionized or protonated to form the initial carbocation intermediates indicated in **(B)**. Branches in **(A)** are colored according to these initial carbocation cyclization mechanisms and the subsequent carbocation formed, shown in **(B)**. Deprotonation of the initial carbocation result in the formation of acyclic sesquiterpenes, as shown in **(B)**.

Functional Annotation of the *VviTPS-g* Subfamily

It was previously shown that the *VviTPS-g* family is expanded in grapevine, forming three distinct clades that separate according to the product profiles of *in vitro* characterized TPS-g enzymes (Martin et al., 2010). Those that accept only GPP as substrate to form geraniol formed a distinct clade with the multi-substrate enzymes forming the other two clades, as shown in **Figures 7A,B**. Annotation of this subfamily to a large extent resolved the current lack in chromosome mapping of this family (**Figure 2**, **Supplementary Figure 2** and **Supplementary Data Sheet 3**). Despite the various improvements of the reference genome, chr.

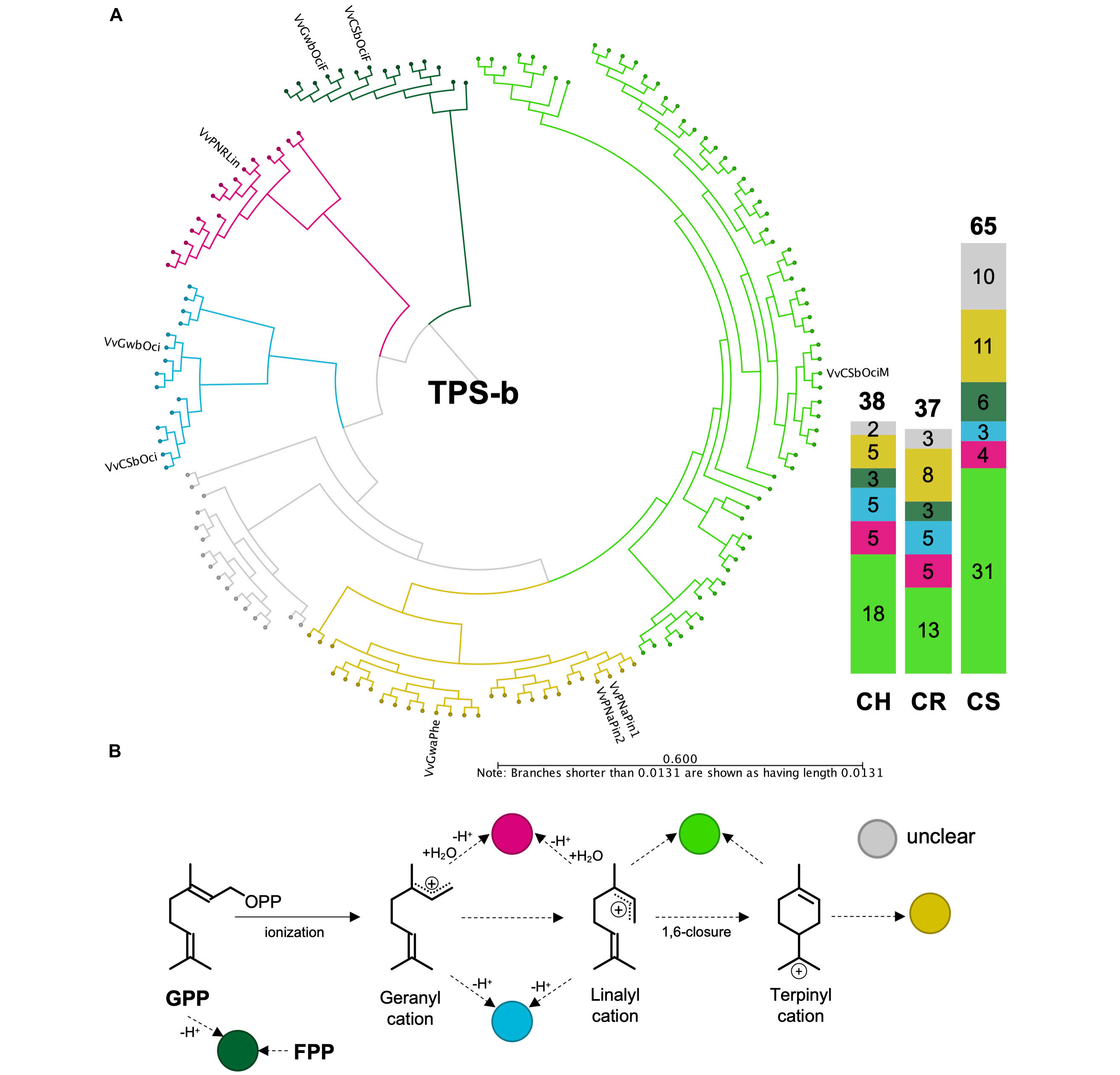


FIGURE 6 | (A) Active site phylogeny of putative VviTPS-b proteins from the diploid genomes and functionally characterized enzymes (outer labels) were used to predict the carbocation cascades that proceed from geranyl diphosphate (GPP), as shown in **(B)**. The dark green branches represent enzymes that can utilize both GPP and farnesyl diphosphate (FPP), functioning as single product enzymes to form the monoterpene ocimene or the sesquiterpene (*E,E*)-alpha-farnesene, respectively. The pink branch is represented by a single product enzyme that proceeds through either the geranyl and/or linalyl cations, with a concerted protonation and deprotonation cascade to form the acyclic monoterpene alcohol limonene. Deprotonation of the geranyl and/or linalyl cations will result in the formation of the acyclic monoterpene ocimene, shown by the blue branch. The light green branch is represented by a multiproduct monoterpene synthase with a cascade proceeding from the linalyl cation to produce acyclic terpenes with a 1,6 ring-closure required to form the terpinyl carbocation and subsequent cyclic terpenes. The yellow branch represents multiproduct enzymes that produce only cyclic monoterpenes, with the gray branch representing a clade without a functional enzyme.

10 remained difficult to assemble, with inbreeding of PN40024 not being able to reduce the extent of heterozygosity. The lack of sufficient resolution for this chromosome therefore resulted in highly discontiguous mapping of diploid contigs to PN40024 chr. 10. This discontiguity resulted in only a single *TPS-g* member being represented on chr. 10 of PN40024 (Martin et al., 2010), with the remaining members being unplaced (i.e., chr. 00). The results presented for the draft diploid genomes therefore provide new chromosome specific information of the *VviTPS-g* subfamily.

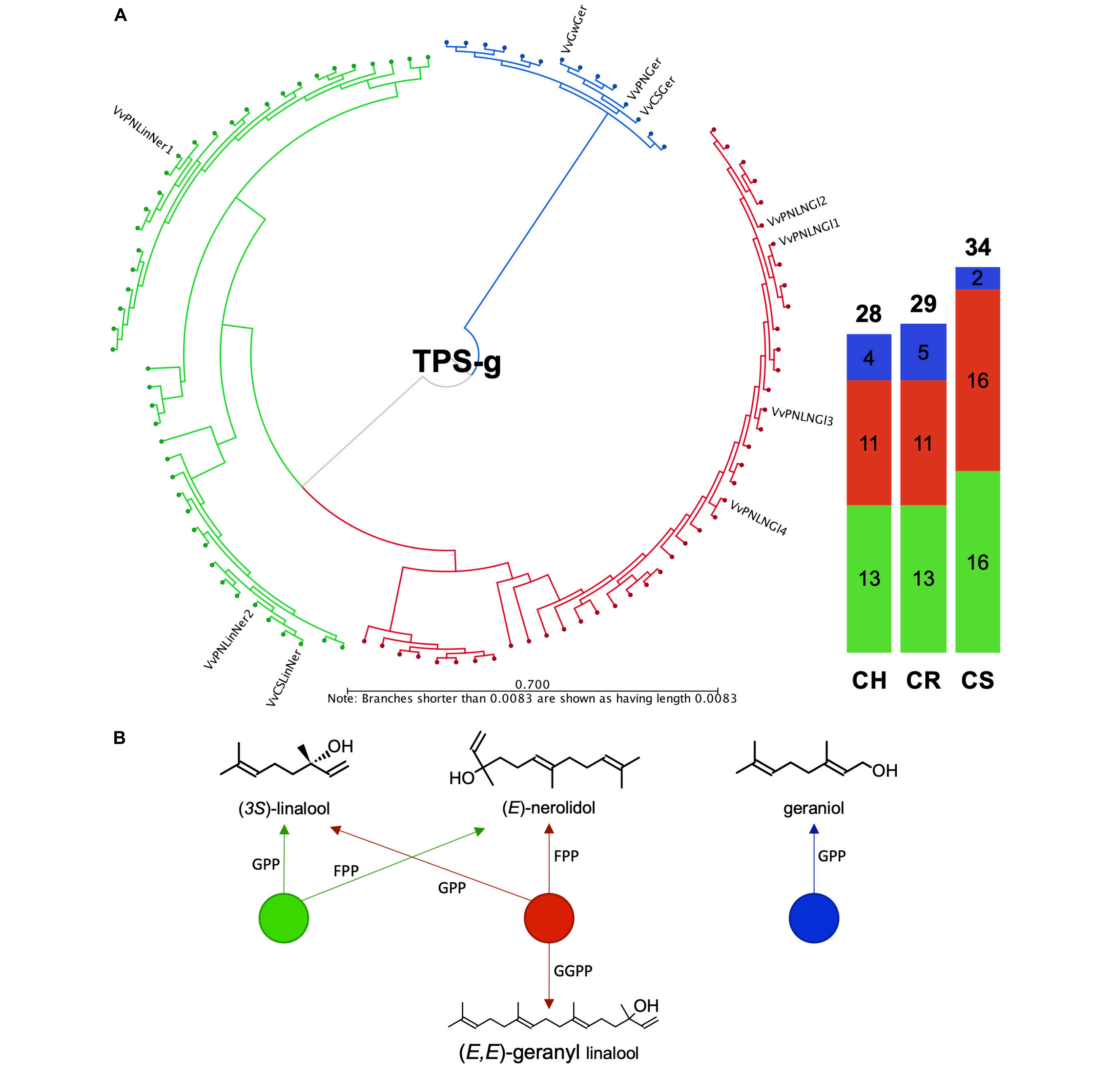


FIGURE 7 | (A) Full length amino acid sequence phylogeny of putative VviTPS-g proteins identified for the diploid genomes and functionally characterized VviTPS-g members (outer label). Branches reflect the enzyme product profile and substrate specificity shown in **(B)**. The green clade consists of single product enzymes that utilize both geranyl diphosphate (GPP) and farnesyl diphosphate (FPP) to produce the acyclic monoterpene alcohol (3*S*)-linalool or acyclic sesquiterpene alcohol (*E*)-nerolidol, respectively. The red clade represents enzymes that in addition to the mechanism shown in green, accept geranyl geranyl diphosphate (GGPP) to produce the acyclic diterpene alcohol (*E*,*E*)-geranyl linalool. The blue clade represents enzymes that only accept GPP to form the acyclic monoterpene alcohol geraniol.

CS had 28/34 *VviTPS-g* members that mapped to chr. 10, of which 14 were located in a 262 kb region of the primary contig VvCabSauv08_v1_Primary000201F. Seven of the genes in this cluster were predicted to be functional and are highly connected to genes from seven different haplotigs, all mapping to chr. 10 (**Figures 4B**, **7**). Furthermore, the *VviTPS* gene order of the primary contig was dissimilar to the haplotigs with large size differences for the intergenic regions, indicating a high level of heterozygosity for this chromosome (**Figure 2**, **Supplementary Figure 2**, and **Supplementary Data Sheet 3**). CR had a similar sized *VviTPS-g* family on chr. 10, localizing to two different primary contigs with almost no contiguity to PN-chr. 10 (**Figure 2**, **Supplementary Figure 2**, and **Supplementary Data Sheet 3**). It was evident from the haplotig to primary contig mappings that chr. 10 is also highly heterozygous for CR. CH, was the exception with 17/28 *VviTPS-g* members mapping to chr. 10 (**Figures 2**, **7**, **Supplementary Figure 2**, and **Supplementary Data Sheet 3**). All seventeen are located on a single contig, connected as tandem duplications in **Figure 4C**, suggesting that it is more homozygous for *VviTPS-g* members on chr. 10. As with the other two genomes, this region was highly discontiguous to the reference genome (**Supplementary Data Sheet 3**).

Comparative Genomics Using Interactive Networks

To understand the complexity of the *VviTPS* family, an integrated view of all the components that influence the different subfamilies is required. The network in **Figure 8** shows the *VviTPS* containing chromosomes, gene duplications and putative proteins for the three diploid genomes. Contig nodes were excluded from the visualization but can be accessed in the interactive network online. The three major *VviTPS*-containing chromosomes, namely chr. 13, −18 and −19 show extensive duplications on the respective chromosomes with few shared between chromosomes. Although the remaining chromosomes, excluding chr. 10, have few *VviTPS* genes, it is evident that they are extensively connected between chromosomes, specifically the multi-substrate *TPS-g* family of chr. 10.

An all-against-all clustering of diploid genome putative VviTPS proteins and functionally characterized proteins is shown in **Figure 9**. The network consists of 533 proteins of which 44 are functionally characterized (Lücker et al., 2004; Martin et al., 2009, 2010; Drew et al., 2015; Smit et al., 2019), sized and shaded in **Figure 9**. To date no *VviTPS-c* or *-e* members have been characterized, therefore the three predicted PN40024 members from the respective subfamilies were included as representatives (Martin et al., 2010). The 533 proteins could be clustered into 111 representative sequences (**Supplementary Data Sheet 4**), indicted by the triangular nodes. Of the representative sequences, 24 *VviTPS-a*, 16 *VviTPS-b* and 7 *VviTPS-g* sequences were not connected to any other sequence indicating that they are unique.

The aforementioned results provides and overview of what is available in the respective networks, however, the data generated in this study is intended to be accessed and mined interactively. Networks can be accessed and downloaded through NDEx (Pratt et al., 2015): http://www.ndexbio.org/#/networkset/b90de24a-24fa-11ea-bb65-0ac135e8bacf?accesskey=c3cdbc1558016990ab78cab2e33cdc41b43c8333ea02799413ebb48f58abbe45. **Supplementary Data Sheet 4** contains the representative VviTPS protein sequences, illustrated in **Figure 9**, and allow for the BLAST lookup for genes of interest using the "align two or more sequences" function of protein BLAST[3]. All nodes and edges in the respective networks are clickable and represent the entire collection of the data generated in this study. By interacting with the nodes and edges, a user can find the nearest functionally characterized protein, which includes metadata for NCBI accessions and nearest reference gene model, as predicted by Martin et al. (2010), as well as subfamily specific reaction mechanisms. It is therefore possible to query any new gene of interest against the current *VviTPS* gene family for the three diploid genomes and the PN40024 reference genome. We recommend viewing the networks on a local machine using Cytoscape (Shannon et al., 2003). A help document to guide users through this is made available with in **Supplementary Data Sheet 4**. The curated genomic, coding and protein sequences represented in the various networks are available as FASTA files in **Supplementary Data Sheet 5**.

DISCUSSION

The reference genome highlighted the extensive duplications and functional diversification of the *VviTPS* family (Martin et al., 2010). Although it was stated that paralogous genes are spread across the genome; identifying homologs were not possible due to the cultivar clone sequenced being near-homozygous (Jaillon et al., 2007). A typical approach to find paralogs entails a BLAST search to find a gene of interest, followed by locating it on the genome, usually through a genome browser. Although specialized gene families have been annotated for grapevine (Martin et al., 2010; Vannozzi et al., 2012), delayed incorporation of these annotations into the reference annotation (Canaguier et al., 2017) and limited visibility of these curations often result in an outdated annotation, most commonly 12x.v0, being used to interpret newly generated results (Grimplet and Cramer, 2019). For example, the web interface of Ensembl Plants (Howe et al., 2020) presents the most complete set of tools to analyze the grapevine genome, but still relies on the 12x.v0 assembly and annotation, limiting its use for the analysis of specialized gene families. Furthermore, the Nimblegen microarray platform utilized for numerous grapevine expression studies showed extensive probe ambiguities within the VviTPS family when using the 12x.v0 annotation, misrepresenting the expression patterns of *VviTPS* genes (Smit et al., 2019). The mapping of RNAseq reads to the aforementioned annotation presents a similar challenge, however, *de novo* assemblies of reads allow for more accurate profiling of *VviTPS* expression patterns (Da Silva et al., 2013; Venturini et al., 2013).

The link between *VviTPS* expression patterns and observed metabolites is therefore tenuous, requiring a critical re-evaluation. As we progress into a new generation of highly contiguous phased diploid genomes it is critical for expanded gene families involved in specialized metabolism to be accurately annotated. This is not only important from a wine aroma perspective but also from an ecophysiological perspective. Numerous terpenoids have been shown to provide important fitness advantages; this includes plant defense, abiotic and biotic stress and chemical signaling (reviewed in Pichersky and Raguso, 2018). The latter aspects will become increasingly important as we aim to breed hardier grapevines, with increased tolerance to climate fluctuations while maintaining sought after aromatic qualities.

The approach presented in this study was akin to that of a pangenome but utilizes a network for data visualization rather than a genome browser. Pangenomes typically focus on the differences and similarities between species, however, the

[3] https://blast.ncbi.nlm.nih.gov/Blast.cgi

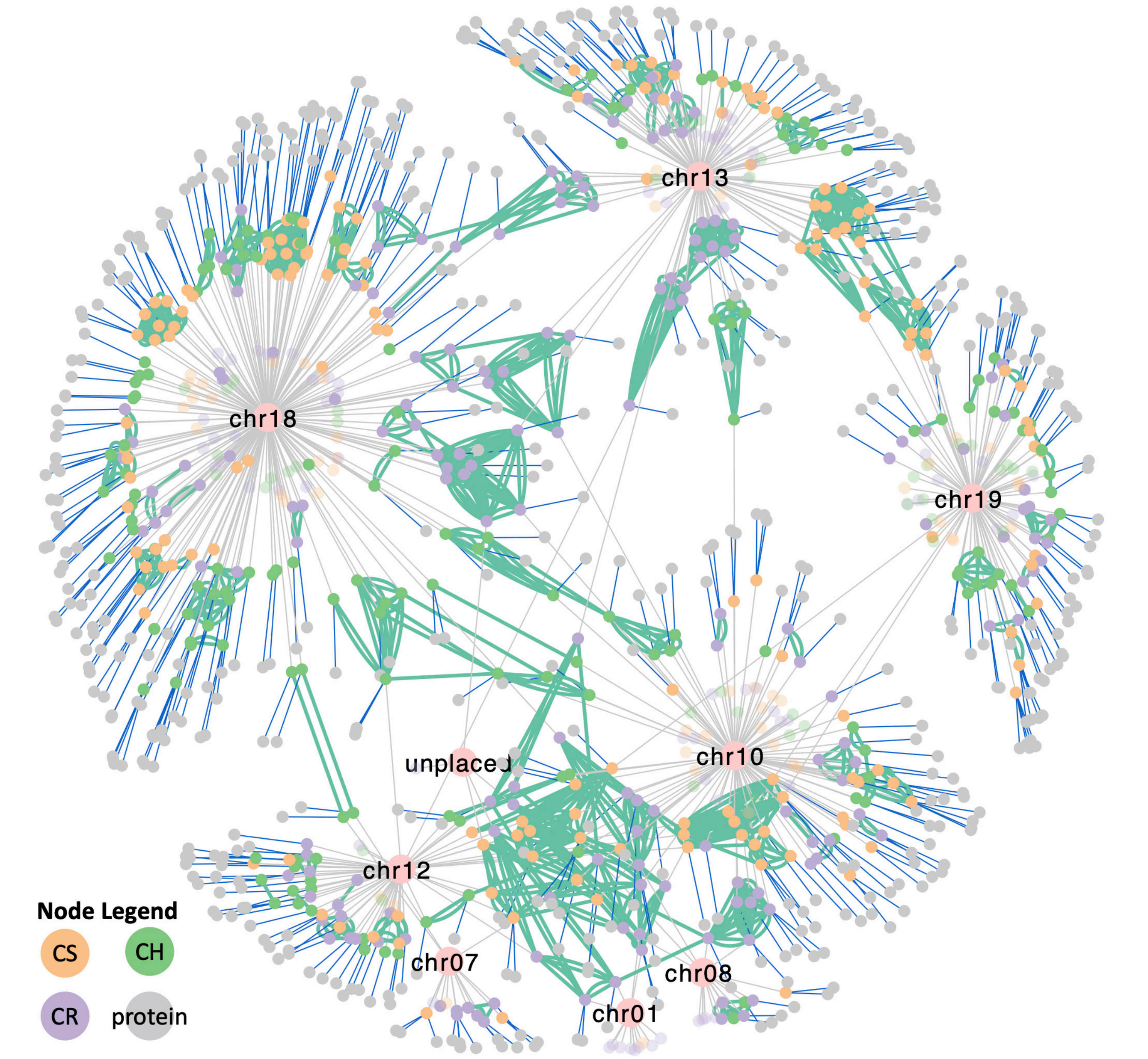


FIGURE 8 | Cytoscape network representing all metadata represented in **Figures 2–7** of the respective *VviTPS* gene families for Cabernet Sauvignon (CS), Carménère (CR), and Chardonnay (CH). Nodes represent *VviTPS* genes, colored according to cultivar with the lighter shade showing partial genes. Gray edges show the chromosome mapping (pink nodes), with green edges showing genes with high homology (degree of similarity (*I'*) > 80%). Green edges connecting nodes of the same color reflect within cultivar homologs (i.e., duplicated genes) while edge connections between different colored nodes show cross-cultivar homologs. Blue edges show putative proteins associated with a gene (i.e., predicted to be functional). An interactive version of this network to access the metadata is available online.

genotypes presented here were expected to be highly similar due to it being closely related cultivars of the same species (**Supplementary Figure 1**). Although partial gene duplications were annotated (refer to the network illustrated by **Figure 8**), their evolutionary importance was not explored further. For the same reason the transposable elements proximal to *VviTPS* genes were excluded. Both of these aspects will become more relevant once the draft diploid genomes are assembled to chromosomes, allowing for in-depth analysis of collinearity and synteny. Nevertheless, the current unassembled genomes allow for a comparative analysis of the *VviTPS* family. The absolute position of *VviTPS* genes on the diploid genomes was therefore not a focus of this study, but rather how genes are related and how their putative function will impact the genetic potential of a genotype. This was possible due to the size of the highly contiguous diploid genome contigs

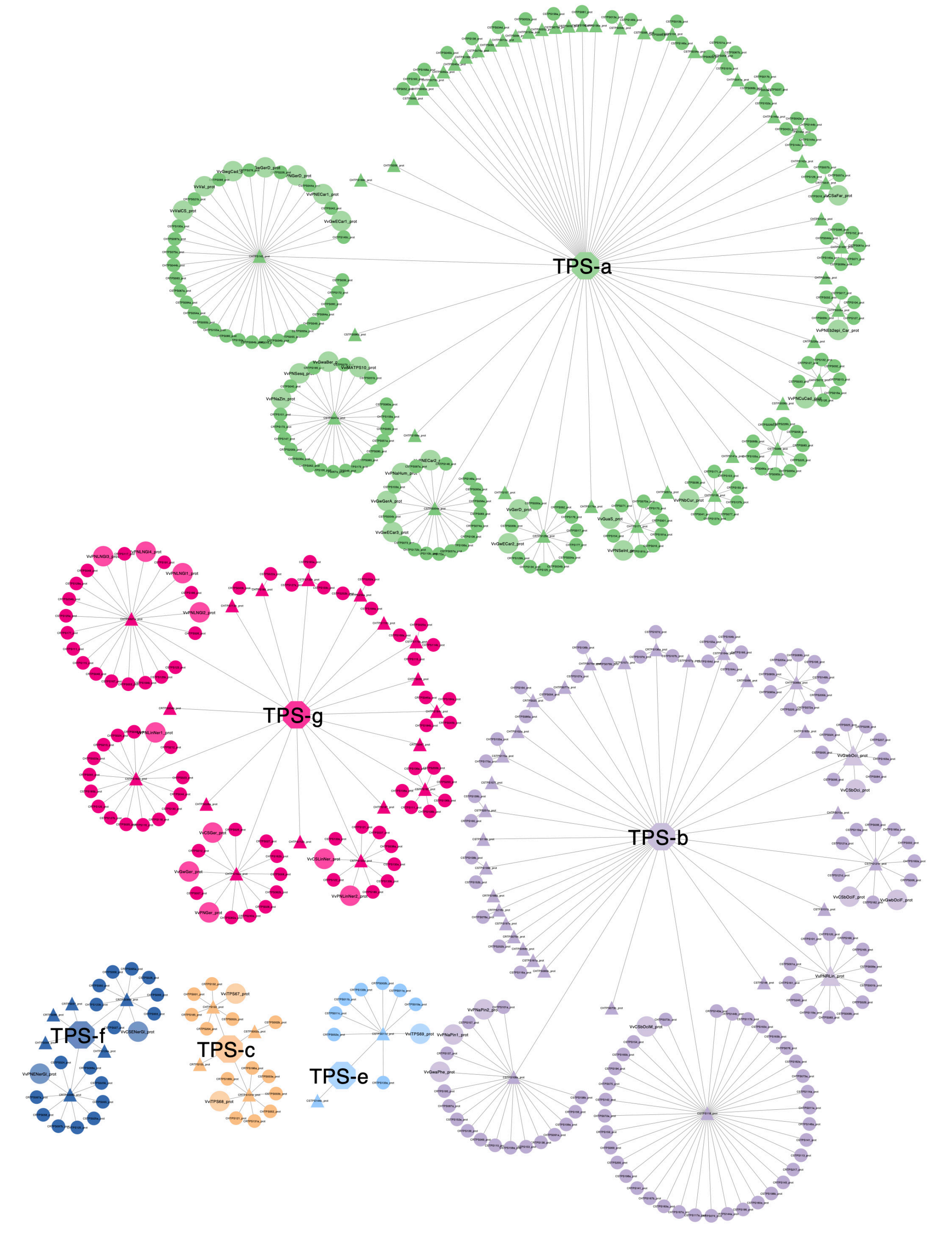


FIGURE 9 | Representative proteins (triangles) show clustering of conserved proteins per *VviTPS* subfamily (central nodes with subfamily name). An absence of connections indicate that the sequence is unique. Enlarged circular nodes are functionally characterized enzymes. The representative sequences serve as target sequences through BLASTp in order to query any TPS of interest to identify mechanistically conserved enzyme clusters. To access the metadata through BLASTp refer to the network available online and the accompanying help document, as described in the supplementary data.

having little to no overlap, in essence each representing a unique genomic region.

Function inference, is however, not purely based on sequence similarity due to the complexity of the carbocation cascades involved in enzyme catalytic mechanisms. The availability of various TPS crystal structures (Lesburg, 1997; Starks, 1997; Williams et al., 1998; Caruthers et al., 2000; Rynkiewicz et al., 2001; Shishova et al., 2007; Gennadios et al., 2009; Li et al., 2013) and functionally characterized enzymes, combined with quantum mechanical modeling (Bülow et al., 2000; Davis and Croteau, 2000; Gao et al., 2012; Miller and Allemann, 2012; Hong and Tantillo, 2014; Wedler et al., 2015; O'Brien et al., 2016; Durairaj et al., 2019), have contributed to elucidating how these cascades proceed in producing the thousands of naturally occurring terpene structures (Osbourn and Lanzotti, 2009; Buckingham et al., 2015).

Sequence identity and protein structure homology to experimentally characterized enzymes have been shown to be an effective approach to predict TPS reaction mechanisms (Degenhardt et al., 2009; Durairaj et al., 2019; Smit et al., 2019). However, this approach requires an extensive understanding of TPS reaction mechanisms, which is especially relevant when considering that the presence of a transcript does not necessarily correspond to a functional enzyme; and that enzyme mechanics can differ (within and between genotypes) due to mutations (Drew et al., 2015; Smit et al., 2019). The results generated in this study therefore provide a multi-genotype view of the *VviTPS* family, consisting of both gene annotation and functional predictions to disseminate and significantly expand on existing knowledge. The benefits of long-read sequencing, allowing for haplotype resolution, despite being unassembled, must be emphasized as it overcomes erroneous assembly of highly similar and duplicated gene regions that could not be resolved through short-read sequencing. The collection of interactive networks therefore provides a platform for studying this family in different grapevine genotypes and provides a novel approach for studying expanded gene families involved in specialized metabolism.

The Grapevine *TPS-g* Family

Mapping of diploid contigs to the reference genome resulted in the identification of *VviTPS-g* members that localize to chr. 10. Two compounding factors made analyzing this family on the reference genome challenging: (1) chr. 10 is known to be highly discontiguous for the reference genome and (2); the PN40024 members of the *VviTPS-g* family are not placed on a chromosome, instead mapping to the chr. 00 pseudo-molecule (Martin et al., 2010; Canaguier et al., 2017). Analysis of the diploid genomes revealed that the contigs localizing to PN40024 chr. 10 had low RaGOO scores and high levels of heterozygosity (**Supplementary Table 2** and **Supplementary Data Sheet 3**). This explains, to a large degree, the discontiguity of chr. 10 and the lack of VviTPS annotations. It may therefore be worthwhile (for the grapevine community) to consider remapping of the PN40024 short-reads to the phased diploid genomes in order to obtain a more contiguous chr. 10 for the reference genome. Nevertheless, the contiguity and size of the phased diploid contigs allowed us to overcome the aforementioned limitations, providing new insights into this important *VviTPS* subfamily (terpene alcohol biosynthesis) and its putative chromosome position.

The diploid genomes, as expected, show an increased number of putative *VviTPS-g* members (28–34 genes) with the function-specific clades being fairly conserved in gene number across the three genomes (**Figure 7**). The phylogenetic distribution within this subfamily, furthermore, highlights the limited number of functionally characterized enzymes that could be used to infer those that potentially contribute to the biosynthesis of terpene alcohols. Of the ten characterized *VviTPS-g* members, seven were characterized from Pinot Noir. Functional groupings in **Figure 7** shows that dual substrate (GPP and FPP) enzymes capable of producing both linalool and nerolidol are overrepresented in all cultivars. Although a large clade of enzymes are predicted to use GGPP as well, resulting in (*E,E*)-geranyl linalool biosynthesis, the ability to use all three substrates *in planta* has not been reported. Subcellular compartmentalization of precursor pools (IPP and DMAPP) and regulation of prenyl substrate biosynthesis is tightly regulated, resulting in compartment-specific biosynthesis of terpenes (Wu et al., 2006; Heinig et al., 2013). Substrate specificity is thought to be affected by the active site, resulting in differential affinities to GPP, FPP and GGPP when enzymes are studied *in vitro* (Arimura et al., 2007; Pazouki et al., 2015). This was also shown for *PlTPS2* from *Phaseolus lunatus* (lima bean), however, *in planta* expression of this gene resulted in (*E,E*)-geranyl linalool and hemiterpene accumulation (Brillada et al., 2013). It is thus likely that the tri-substrate VviTPS-g clade is involved in (*E,E*)-geranyl linalool biosynthesis rather than (*3S*)-linalool and/or (*E*)-nerolidol biosynthesis.

The clade for geraniol biosynthesis had only two putatively functional proteins for CS, with CH and CR having 4 and 5, respectively (**Figure 7**). During winemaking, geraniol is readily metabolized by yeast during fermentation to form important wine odorants that, along with nerolidol and linalool derivatives, make up the core constituents of aromatic wines, often described as having a Muscat or "floral" aromas (King and Dickinson, 2000; Emanuelli et al., 2010). These transformations are facilitated by specific yeast genera that facilitate the reduction of the terpenoid or cleavage of glycosyl groups. The available substrate (cultivar-specific terpenoids) and vinification style will therefore directly influence the extent of floral aroma catalysis (Carrau et al., 2005; Cramer et al., 2014).

Furthermore, (*E*)-nerolidol and (*E,E*)-geranyl linalool are known precursors for the homoterpenes (E)-4,8-dimethyl-1,3,7-nonatriene (DMNT) and (E,E)- 4,8,12-trimethyltrideca-1,3,7,11-tetraene (TMTT), respectively. DMNT, is especially important from an ecological perspective due to it being emitted by various grapevine organs, with flower and leaf emissions linked to the attraction of the grapevine berry moth, *Lobesia botrana,* a major grapevine pest (Tasin et al., 2007). Recent efforts to alter the chemical emission profile of grapevine focused on overexpressing an (*E*)-beta-farnesene synthase, decreasing *L. botrana* attraction to grapevine (Salvagnin et al., 2018). The numerous *TPS-g* members annotated here therefore provide alternative targets to alter (*E,E*)-geranyl linalool, and by extension DMNT, biosynthesis.

The *VviTPS-a* and *-b* Subfamilies: An Expanded Group With Specialized Reaction Mechanisms

The TPS-a and -b subfamilies are hypothesized to have evolved from diterpene synthases where the loss of the γ domain or transit peptide, coupled with changes in the active site, lead to neofunctionalization (Köksal et al., 2011a,b; Pazouki and Niinemets, 2016). This likely allowed for spatial-temporal regulation and specialization with vestigial functions explaining the ability to use multiple substrates *in vitro* (Pazouki and Niinemets, 2016).

Sesquiterpene synthases (VviTPS-a) represent the largest grapevine subfamily and are of special interest due to their ability to produce either a single terpene or a multitude of compounds. The diversity in sesquiterpenes is largely due to the extra double bond in FPP, compared to GPP, with subsequent isomerization to NPP resulting in further diversity. Currently accepted reaction mechanisms of plant sesquiterpene synthases (Durairaj et al., 2019) resulted in VviTPS-a members grouping according to which of these are used initial substrate (**Figure 5**). Premature quenching of the cyclization reaction, regardless of whether FPP or NPP is the initial substrate, results in the formation of acyclic sesquiterpenes, with two small but distinct clades (**Figure 5**) suggesting that there may be a distinction in substrate affinity. The isomerization step is rate-limiting (Cane et al., 1997; Miller and Allemann, 2012) which could explain why fewer enzymes are in the NPP clade, suggesting a possible specialized *in planta* function. It was previously shown that PN40024 had distinct clades for 1,10 and 1,11-cyclizations of FPP (Martin et al., 2010; Smit et al., 2019), however, the increased number of putative VviTPS-a proteins from the three diploid genomes added greater complexity to the conservation of enzyme mechanisms (**Figures 5**, **9**). Three distinct clades were identified (**Figure 5**) with the functional enzymes of clades 2 and 3 sharing the same product profiles with a clear distinction between 1,10- and 1,11-cyclizations. The 1,10-cyclizations will proceed through the (*E,E*)-germacradienyl cation to either germacrene A or D as reactive intermediates. From the germacrene A intermediate, an alkyl migration of the eudesmyl cation will be necessary to explain the mechanism for enzymes in clade 2 of **Figure 5A**. A lack of such a migration and the presence of selinene-type synthases are congruent with the reaction mechanisms of enzymes in clade 1 of **Figure 5A** (Caruthers et al., 2000; Calvert et al., 2002; Christianson, 2017). Due to these subtle complexities in enzyme mechanics, VviTPS-a functional predictions is limited to initial substrate and first cyclization (**Figure 5**).

The PN40024 *VviTPS-b* subfamily consists of 45 loci, including pseudogenes, of which 19 were predicted to be functional. Seven of the nineteen have been functionally characterized, resulting in nine novel enzymes. The three phased diploid genomes contain between 37 and 65 *VviTPS-b* complete genes (fl- and d-ORF), excluding partial genes (**Figure 6**), providing an extended number of new *VviTPS-b* gene models. Although multiple reaction mechanism was identified for clades within the TPS-b subfamily, the overarching differences were between TPS-b Type I and II mechanisms. It was, however, noted that the single product Type II enzymes formed unique clades. This was also reported by Martin et al. (2010) where the two reaction types were bifurcated by sequences from other plants instead of a group of enzymes with no clear function, shown in **Figure 6**. The clades in **Figure 6**, indicate a conserved set of Type II enzymes that seemingly evolved to multi-product Type I enzymes. The clade of proteins associated with enzymes that accept FPP *in vitro* seems to be conserved, with its phylogenetic position supporting the specialization hypothesis (Pazouki and Niinemets, 2016).

CONCLUSION

The availability of new genomic resources allowed for a comparative analysis of the *VviTPS* family, expanding on what the PN40024 genome offered. The resolution of haplotypes allowed for the identification of putative alleles with greater sequence contiguity, due to long-read sequencing, allowing for a comprehensive, and more complete annotation of this expanded gene family. Phylogenomic similarity and functional predictions greatly benefited from having expanded genotypic variation. This allowed for greater subfamily-specific functional predictions while addressing specific limitations on the reference genome, particularly the *VviTPS-g* subfamily. The data presented is not intended to be a static resource with the incorporation of inter-varietal and -species variations at genomic and single base-pair levels expected to improve the accuracy of functional predictions. The recent release of a phased *V. riparia* genome (Girollet et al., 2019) and nucleotide variation data from 472 *Vitis* species (Liang et al., 2019), specifically hold great promise for elucidating the impact that domestication and breeding had on *VviTPS* evolution, expansion and functionalization. Although the diploid genomes are currently available as draft assemblies, this limitation is expected to be addressed in the near future. Establishing congruency with the reference genome will most likely require a critical re-evaluation of the PN40024 genome assembly to address the numerous limitations regarding its completeness and contiguity. The utilization of networks to show relatedness of *VviTPS* genes at the genomic, coding and protein sequence levels within and between cultivars provides a novel, valuable and interactive resource. This resource is intended to provide a starting platform from which genotypic variation can be explored and expanded on to characterize the *VviTPS* family further, while providing a blueprint for future comparative analyses of specialized gene families.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NDEx repository http://www.ndexbio.org/#/networkset/b90de24a-24fa-11ea-bb65-0ac135e8bacf?accesskey=c3cdbc1558016990ab78cab2e33cdc41b43c8333ea02799413ebb48f58abbe45.

AUTHOR CONTRIBUTIONS

SS, MV, and PY conceptualized the study. SS performed all computational analyses, gene annotations and drafted the initial manuscript. All authors contributed to the final manuscript.

FUNDING

The study was financially supported with grants from Wine Industry Network for Expertise and Technology (Winetech; IWBT P14/02), the National Research Foundation (NRF) Thuthuka (TTK13070220277), and the Technology and Human Resources for Industry Programme (THRIP) of the Department of Trade and Industry (DTI).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00421/full#supplementary-material

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., et al. (2019). Fast and accurate reference-guided scaffolding of draft genomes. *bioRxiv* [Preprint]. doi: 10.1101/519637

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Arimura, G., Garms, S., Maffei, M., Bossi, S., Schulze, B., Leitner, M., et al. (2007). Herbivore-induced terpenoid emission in *Medicago truncatula*: concerted action of jasmonate, ethylene and calcium signaling. *Planta* 227, 453–464. doi: 10.1007/s00425-007-0631-y

Bick, J. A., and Lange, B. M. (2003). Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. *Arch. Biochem. Biophys.* 415, 146–154. doi: 10.1016/S0003-9861(03)00233-9

Black, C. A., Parker, M., Siebert, T. E., Capone, D. L., and Francis, I. L. (2015). Terpenoids and their role in wine flavour: recent advances. *Aust. J. Grape Wine Res.* 21, 582–600. doi: 10.1111/ajgw.12186

Bloch, K., Chaykin, S., Phillips, A., and De Waard, A. (1959). Mevalonic acid pyrophospahte and isopentenylpyrphosphate. *J. Biol. Chem.* 234, 2595–2604.

Bohlmann, J., Meyer-Gauen, G., and Croteau, R. (1998). Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4126–4133. doi: 10.1073/pnas.95.8.4126

Brillada, C., Nishihara, M., Shimoda, T., Garms, S., Boland, W., Maffei, M. E., et al. (2013). Metabolic engineering of the C16 homoterpene TMTT in lotus japonicus through overexpression of (E,E)-geranyllinalool synthase attracts generalist and specialist predators in different manners. *New Phytol.* 200, 1200–1211. doi: 10.1111/nph.12442

Buckingham, J., Cooper, C. M., and Purchase, R. (2015). *Natural Products Desk Reference*. Boca Raton, FL: CRC Press.

Bülow, N., König, W., and Konig, W. (2000). The role of germacrene D as a precursor in sesquiterpene biosynthesis: investigations of acid catalyzed, photochemically and thermally induced rearrangements. *Phytochemistry* 55, 141–168. doi: 10.1016/s0031-9422(00)00266-1

Calvert, M. J., Ashton, P. R., and Allemann, R. K. (2002). Germacrene A is a product of the aristolochene synthase-mediated conversion of farnesylpyrophosphate to aristolochene. *J. Am. Chem. Soc.* 124, 11636–11641. doi: 10.1021/ja020762p

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Canaguier, A., Grimplet, J., Di Gaspero, G., Scalabrin, S., Duchêne, E., and Choisne, N. (2017). A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). *Genomics Data* 14, 56–62. doi: 10.1016/J.GDATA.2017.09.002

Cane, D. E. (1990). Enzymic formation of sesquiterpenes. *Chem. Rev.* 90, 1089–1103. doi: 10.1021/cr00105a002

Cane, D. E., Chiu, H. T., Liang, P. H., and Anderson, K. S. (1997). Pre-steady-state kinetic analysis of the trichodiene synthase reaction pathway. *Biochemistry* 36, 8332–8339. doi: 10.1021/bi963018o

Carrau, F. M., Medina, K., Boido, E., Farina, L., Gaggero, C., Dellacassa, E., et al. (2005). *De novo* synthesis of monoterpenes by *Saccharomyces cerevisiae* wine yeasts. *FEMS Microbiol. Lett.* 243, 107–115. doi: 10.1016/j.femsle.2004.11.050

Caruthers, J. M., Kang, I., Rynkiewicz, M. J., Cane, D. E., and Christianson, D. W. (2000). Crystal structure determination of aristolochene synthase from the blue cheese mold, penicillium roqueforti. *J. Biol. Chem.* 275, 25533–25539. doi: 10.1074/jbc.M000433200

Chin, C., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi: 10.1038/nmeth.4035

Christianson, D. W. (2006). Structural biology and chemistry of the terpenoid cyclases. *Chem. Rev.* 106, 3412–3442. doi: 10.1021/cr050286w

Christianson, D. W. (2017). Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* 117, 11570–11648. doi: 10.1021/acs.chemrev.7b00287

Cramer, G. R., Ghan, R., Schlauch, K. A., Tillett, R. L., Heymann, H., Ferrarini, A., et al. (2014). Transcriptomic analysis of the late stages of grapevine (*Vitis vinifera* cv. cabernet sauvignon) berry ripening reveals significant induction of ethylene signaling and flavor pathways in the skin. *BMC Plant Biol.* 14:370. doi: 10.1186/s12870-014-0370-8

Da Silva, C., Zamperin, G., Ferrarini, A., Minio, A., Dal Molin, A., Venturini, L., et al. (2013). The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* 25, 4777–4788. doi: 10.1105/tpc.113.118810

Davis, E. M., and Croteau, R. (2000). Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. *Top. Curr. Chem.* 209, 53–95. doi: 10.1007/3-540-48146-x_2

Degenhardt, J., Köllner, T. G., and Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* 70, 1621–1637. doi: 10.1016/j.phytochem.2009.07.030

Dèrozier, S., Samson, F., Tamby, J.-P., Guichard, C., Brunaud, V., Grevet, P., et al. (2011). Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods* 7:8. doi: 10.1186/1746-4811-7-8

D'Onofrio, C., Matarese, F., and Cuzzola, A. (2017). Study of the terpene profile at harvest and during berry development of *Vitis vinifera* L. aromatic varieties Aleatico, Brachetto, Malvasia di Candia aromatica and Moscato bianco. *J. Sci. Food Agric.* 97, 2898–2907. doi: 10.1002/jsfa.8126

Drew, D. P., Andersen, T. B., Sweetman, C., Møller, B. L., Ford, C., and Simonsen, H. T. (2015). Two key polymorphisms in a newly discovered allele of the *Vitis vinifera* TPS24 gene are responsible for the production of the rotundone precursor α-guaiene. *J. Exp. Bot.* 67, 799–808. doi: 10.1093/jxb/erv491

Dueholm, B., Drew, D. P., Sweetman, C., and Simonsen, H. T. (2019). In planta and in silico characterization of five sesquiterpene synthases from *Vitis vinifera* (cv. Shiraz) berries. *Planta* 249, 59–70. doi: 10.1007/s00425-018-2986-7

Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., and van Dijk, A. D. (2019). An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* 158, 157–165. doi: 10.1016/j.phytochem.2018.10.020

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Emanuelli, F., Battilana, J., Costantini, L., Le Cunff, L., Boursiquot, J.-M., This, P., et al. (2010). A candidate gene association study on muscat flavor in grapevine (*Vitis vinifera* L.). *BMC Plant Biol.* 10:241. doi: 10.1186/1471-2229-10-241

Figueroa-Balderas, R., Minio, A., Morales-Cruz, A., Vondras, A. M., and Cantu, D. (2019). "Strategies for sequencing and assembling grapevine genomes," in

The Grape Genome, eds D. Cantu and M. A. Walker (Cham: Springer), 77–88. doi: 10.1007/978-3-030-18601-2_5

Gao, Y., Honzatko, R. B., and Peters, R. J. (2012). Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat. Prod. Rep.* 29, 1153–1175. doi: 10.1039/c2np20059g

Gennadios, H. A., Gonzalez, V., Di Costanzo, L., Li, A., Yu, F., Miller, D. J., et al. (2009). Crystal structure of (+)-δ-cadinene synthase from *Gossypium arboreum* and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry* 48, 6175–6183. doi: 10.1021/bi900483b

Girollet, N., Rubio, B., and Bert, P.-F. (2019). *De novo* phased assembly of the *Vitis riparia* grape genome. *Sci. Data* 6:127. doi: 10.1038/s41597-019-0133-3

Grimplet, J., and Cramer, G. R. (2019). "The grapevine genome annotation," in *The Grape Genome*, eds D. Cantu and M. A. Walker (Cham: Springer), 89–101. doi: 10.1007/978-3-030-18601-2_6

Heinig, U., Gutensohn, M., Dudareva, N., and Aharoni, A. (2013). The challenges of cellular compartmentalization in plant metabolic engineering. *Curr. Opin. Biotechnol.* 24, 239–246. doi: 10.1016/j.copbio.2012.11.006

Hemmerlin, A., Hoeffler, J. F., Meyer, O., Tritsch, D., Kagan, I. A., Grosdemange-Billiard, C., et al. (2003). Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *J. Biol. Chem.* 278, 26666–26676. doi: 10.1074/jbc.M302526200

Hjelmeland, A. K., Zweigenbaum, J., and Ebeler, S. E. (2015). Profiling monoterpenol glycoconjugation in *Vitis vinifera* L. cv. Muscat of Alexandria using a novel putative compound database approach, high resolution mass spectrometry and collision induced dissociation fragmentation analysis. *Anal. Chim. Acta* 887, 138–147. doi: 10.1016/j.aca.2015.06.026

Hong, Y. J., and Tantillo, D. J. (2014). Branching out from the bisabolyl cation. unifying mechanistic pathways to barbatene, bazzanene, chamigrene, chamipinene, cumacrene, cuprenene, dunniene, isobazzanene, iso-γ-bisabolene, isochamigrene, laurene, microbiotene, sesquithujene, sesquisabinene, t. *J. Am. Chem. Soc.* 136, 2450–2463. doi: 10.1021/ja4106489

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., et al. (2020). Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48, D689–D695. doi: 10.1093/nar/gkz890

Hyatt, D. C., Youn, B., Zhao, Y., Santhamma, B., Coates, R. M., Croteau, R. B., et al. (2007). Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 5360–5365. doi: 10.1073/pnas.0700915104

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jukes, T. H., and Cantor, C. R. (1969). "Evolution of protein molecules," in *Mammalian Protein Metabolism*, ed. H. N. Munro (New York, NY: Academic Press), 21–132. doi: 10.1016/B978-1-4832-3211-9.50009-7

Kalua, C. M., and Boss, P. K. (2009). Evolution of volatile compounds during the development of cabernet sauvignon grapes (*Vitis vinifera* L.). *J. Agric. Food Chem.* 57, 3818–3830. doi: 10.1021/jf803471n

King, A., and Dickinson, J. R. (2000). Biotransformation of monoterpene alcohols by *Saccharomyces cerevisiae*, *Torulaspora delbrueckii* and *Kluyveromyces lactis*. *Yeast* 16, 499–506. doi: 10.1002/(sici)1097-0061(200004)16:6<499::aid-yea548>3.0.co;2-e

Köksal, M., Hu, H., Coates, R. M., Peters, R. J., and Christianson, D. W. (2011a). Structure and mechanism of the diterpene cyclase ent-copalyl diphosphate synthase. *Nat. Chem. Biol.* 7, 431–433. doi: 10.1038/nchembio.578

Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011b). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* 469, 116–122. doi: 10.1038/nature09628

Larsson, J. (2019). *eulerr : Area-Proportional Euler and Venn Diagrams With Ellipses*. Available online at: https://cran.r-project.org/package=eulerr (accessed November 1, 2019).

Lesburg, C. A. (1997). Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* 277, 1820–1824. doi: 10.1126/science.277.5333.1820

Li, J.-X., Fang, X., Zhao, Q., Ruan, J.-X., Yang, C.-Q., Wang, L.-J., et al. (2013). Rational engineering of plasticity residues of sesquiterpene synthases from *Artemisia annua*?: product specificity and catalytic efficiency. *Biochem. J.* 451, 417–426. doi: 10.1042/BJ20130041

Li, W. H., Gu, Z., Wang, H., and Nekrutenko, A. (2001). Evolutionary analyses of the human genome. *Nature* 409, 847–849. doi: 10.1038/35057039

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., et al. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history. *Nat. Commun.* 10:1190. doi: 10.1038/s41467-019-09135-8

Lichtenthaler, H. K. (1999). The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 47–65. doi: 10.1146/annurev.arplant.50.1.47

Lin, J., Massonnet, M., and Cantu, D. (2019). The genetic basis of grape and wine aroma. *Hortic. Res.* 6:81. doi: 10.1038/s41438-019-0163-1

Lücker, J., Bowen, P., and Bohlmann, J. (2004). *Vitis vinifera* terpenoid cyclases: functional identification of two sesquiterpene synthase cDNAs encoding (+)-valencene synthase and (-)-germacrene D synthase and expression of mono- and sesquiterpene synthases in grapevine flowers and berries. *Phytochemistry* 65, 2649–2659. doi: 10.1016/j.phytochem.2004.08.017

Martin, D., Aubourg, S., Schouwey, M., Daviet, L., Schalk, M., Toub, O., et al. (2010). Functional Annotation, Genome Organization and Phylogeny of the Grapevine (*Vitis vinifera*) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays. *BMC Plant Biol.* 10:226. doi: 10.1186/1471-2229-10-226

Martin, D. M., Toub, O., Chiang, A., Lo, B. C., Ohse, S., Lund, S. T., et al. (2009). The bouquet of grapevine (*Vitis vinifera* L. cv. Cabernet Sauvignon) flowers arises from the biosynthesis of sesquiterpene volatiles in pollen grains. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7245–7250. doi: 10.1073/pnas.0901387106

Martin, S., and Leggett, R. M. (2019). Alvis: a tool for contig and read ALignment VISualisation and chimera detection. *bioRxiv* [Preprint]. doi: 10.1101/663401

Maul, E., and Töpfer, R. (2015). Vitis International Variety Catalogue (V IVC): a cultivar database referenced by genetic profiles and morphology. *BIO Web Conf.* 5:01009. doi: 10.1051/bioconf/20150501009

Miller, D. J., and Allemann, R. K. (2012). Sesquiterpene synthases: passive catalysts or active players? *Nat. Prod. Rep.* 29, 60–71. doi: 10.1039/C1NP00060H

Minio, A., Lin, J., Gaut, B. S., and Cantu, D. (2017). How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* 8:826. doi: 10.3389/fpls.2017.00826

Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A., and Cantu, D. (2019). Diploid genome assembly of the wine grape Carménère. *G3* 9, 1331–1337. doi: 10.1534/g3.119.400030

Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., et al. (2011). Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3530–3535. doi: 10.1073/pnas.1009363108

O'Brien, T. E., Bertolani, S. J., Tantillo, D. J., and Siegel, J. B. (2016). Mechanistically informed predictions of binding modes for carbocation intermediates of a sesquiterpene synthase reaction. *Chem. Sci.* 7, 4009–4015. doi: 10.1039/C6SC00635C

Osbourn, A. E., and Lanzotti, V. (eds) (2009). *Plant-derived Natural Products*. New York, NY: Springer. doi: 10.1007/978-0-387-85498-4.

Pazouki, L., Memari, H. R., Kännaste, A., Bichele, R., and Niinemets, Ü. (2015). Germacrene A synthase in yarrow (*Achillea millefolium*) is an enzyme with mixed substrate specificity: gene cloning, functional characterization and expression analysis. *Front. Plant Sci.* 6:111. doi: 10.3389/fpls.2015.00111

Pazouki, L., and Niinemets, Ü. (2016). Multi-substrate terpene synthases: their occurrence and physiological significance. *Front. Plant Sci.* 7:1019. doi: 10.3389/fpls.2016.01019

Pichersky, E., and Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytol.* 220, 692–702. doi: 10.1111/nph.14178

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., et al. (2015). NDEx, the network data exchange. *Cell Syst.* 1, 302–305. doi: 10.1016/j.cels.2015.10.001

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Available online at: http://www.r-project.org/ (accessed November 1, 2019).

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Roach, M. J., Johnson, D. L., Bohlmann, J., van Vuuren, H. J. J., Jones, S. J. M., Pretorius, I. S., et al. (2018). Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet.* 14:e1007807. doi: 10.1371/journal.pgen.1007807

Robinson, A. L., Boss, P. K., Solomon, P. S., Trengove, R. D., Heymann, H., and Ebeler, S. E. (2014). Origins of grape and wine aroma. Part 1. Chemical components and viticultural impacts. *Am. J. Enol. Vitic.* 65, 1–24. doi: 10.5344/ajev.2013.12070

Rohmer, M. (1999). The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat. Prod. Rep.* 16, 565–574. doi: 10.1039/a709175c

Rynkiewicz, M. J., Cane, D. E., and Christianson, D. W. (2001). Structure of trichodiene synthase from *Fusarium sporotrichioides* provides mechanistic inferences on the terpene cyclization cascade. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13543–13548. doi: 10.1073/pnas.231313098

Rynkiewicz, M. J., Cane, D. E., and Christianson, D. W. (2002). X-ray crystal structures of D100E trichodiene synthase and its pyrophosphate complex reveal the basis for terpene product diversity. *Biochemistry* 41, 1732–1741. doi: 10.1021/bi011960g

Salvagnin, U., Malnoy, M., Thöming, G., Tasin, M., Carlin, S., Martens, S., et al. (2018). Adjusting the scent ratio: using genetically modified *Vitis vinifera* plants to manipulate European grapevine moth behaviour. *Plant Biotechnol. J.* 16, 264–271. doi: 10.1111/pbi.12767

Schwab, W., Williams, D. C., Davis, E. M., and Croteau, R. (2001). Mechanism of monoterpene cyclization: stereochemical aspects of the transformation of noncyclizable substrate analogs by recombinant (-)-limonene synthase, (+)-bornyl diphosphate synthase, and (-)-pinene synthase. *Arch. Biochem. Biophys.* 392, 123–136. doi: 10.1006/abbi.2001.2442

Schwab, W., and Wüst, M. (2015). Understanding the constitutive and induced biosynthesis of mono- and sesquiterpenes in grapes (*Vitis vinifera*) – A key to unlocking the biochemical secrets of unique grape aroma profiles. *J. Agric. Food Chem.* 63, 10591–10603. doi: 10.1021/acs.jafc.5b04398

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shishova, E. Y., Di Costanzo, L., Cane, D. E., and Christianson, D. W. (2007). X-ray crystal structure of aristolochene synthase from *Aspergillus terreus* and evolution of templates for the cyclization of farnesyl diphosphate. *Biochemistry* 46, 1941–1951. doi: 10.1021/bi0622524

Siebert, T. E., Wood, C., Elsey, G. M., and Pollnitz, A. P. (2008). Determination of rotundone, the pepper aroma impact compound, in grapes and wine. *J. Agric. Food Chem.* 56, 3745–3748. doi: 10.1021/jf800184t

Skinkis, P. A., Bordelon, B. P., and Wood, K. V. (2008). Comparison of monoterpene constituents in Traminette, Gewürztraminer, and Riesling winegrapes. . *Am. J. Enol. Vitic.* 59, 440–445.

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi: 10.1186/1471-2105-6-31

Smit, S. J., Vivier, M. A., and Young, P. R. (2019). Linking terpene synthases to sesquiterpene metabolism in grapevine flowers. *Front. Plant Sci.* 10:177. doi: 10.3389/fpls.2019.00177

Starks, C. M. (1997). Structural basis for cyclic terpene biosynthesis by Tobacco 5-Epi-Aristolochene synthase. *Science* 277, 1815–1820. doi: 10.1126/science.277.5333.1815

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988

Swiegers, J. H., Bartowsky, E. J., Henschke, P. A., and Pretorius, I. S. (2005). Yeast and bacterial modulation of wine aroma and flavour. *Aust. J. Grape Wine Res.* 11, 139–173. doi: 10.1111/j.1755-0238.2005.tb00285.x

Tasin, M., Bäckman, A.-C., Coracini, M., Casado, D., Ioriatti, C., and Witzgall, P. (2007). Synergism and redundancy in a plant volatile blend attracting grapevine moth females. *Phytochemistry* 68, 203–209. doi: 10.1016/j.phytochem.2006.10.015

This, P., Lacombe, T., and Thomas, M. R. (2006). Historical origins and genetic diversity of wine grapes. *Trends Genet.* 22, 511–519. doi: 10.1016/j.tig.2006.07.008

Vannozzi, A., Dry, I. B., Fasoli, M., Zenoni, S., and Lucchin, M. (2012). Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.* 12:130. doi: 10.1186/1471-2229-12-130

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* 2:e1326. doi: 10.1371/journal.pone.0001326

Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G. B., Fasoli, M., Dal Santo, S., et al. (2013). *De novo* transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. *BMC Genomics* 14:41. doi: 10.1186/1471-2164-14-41

Wedler, H., Pemberton, R., and Tantillo, D. (2015). Carbocations and the complex flavor and bouquet of wine: mechanistic aspects of terpene biosynthesis in wine grapes. *Molecules* 20, 10781–10792. doi: 10.3390/molecules200610781

Williams, D. C., McGarvey, D. J., Katahira, E. J., and Croteau, R. (1998). Truncation of limonene synthase preprotein provides a fully active "pseudomature" form of this monoterpene cyclase and reveals the function of the amino-terminal arginine pair. *Biochemistry* 37, 12213–12220. doi: 10.1021/bi980854k

Wood, C., Siebert, T. E., Parker, M., Capone, D. L., Elsey, G. M., Pollnitz, A. P., et al. (2008). From wine to pepper: rotundone, an obscure sesquiterpene, is a potent spicy aroma compound. *J. Agric. Food Chem.* 56, 3738–3744. doi: 10.1021/jf800183k

Wu, S., Schalk, M., Clark, A., Miles, R. B., Coates, R., and Chappell, J. (2006). Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. *Nat. Biotechnol.* 24, 1441–1447. doi: 10.1038/nbt1251

Xu, J., Xu, J., Ai, Y., Farid, R. A., Tong, L., and Yang, D. (2018). Mutational analysis and dynamic simulation of S-limonene synthase reveal the importance of Y573: insight into the cyclization mechanism in monoterpene synthases. *Arch. Biochem. Biophys.* 638, 27–34. doi: 10.1016/j.abb.2017.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fgene.2020.00330

Preferential Subgenome Elimination and Chromosomal Structural Changes Occurring in Newly Formed Tetraploid Wheat—*Aegilops ventricosa* Amphiploid ($AABBD^vD^vN^vN^v$)

Jie Zhang[1,2], Fan Yang[3], Yun Jiang[1], Yuanlin Guo[1], Ying Wang[1], XinGuo Zhu[3], Jun Li[2,3], Hongshen Wan[2,3], Qin Wang[3], Ziyuan Deng[1], Pu Xuan[4] and WuYun Yang[2,3]*

[1] *Institute of Biotechnology and Nuclear Technology Research, Sichuan Academy of Agricultural Sciences, Chengdu, China,* [2] *Key Laboratory of Wheat Biology and Genetic Improvement on Southwestern China (Ministry of Agriculture), Chengdu, China,* [3] *Institute of Crop Research, Sichuan Academy of Agricultural Sciences, Chengdu, China,* [4] *Institute of Agro-products Processing Science and Technology, Sichuan Academy of Agricultural Sciences, Chengdu, China*

OPEN ACCESS

Edited by:
Ray Ming,
University of Illinois
at Urbana-Champaign, United States

Reviewed by:
Claudio Oliveira,
São Paulo State University, Brazil
Lenin Arias Rodriguez,
Universidad Juárez Autónoma
de Tabasco, Mexico
Maciej Majka,
Institute of Plant Genetics (PAN),
Poland

Correspondence:
WuYun Yang
yangwuyun@126.com

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *06 September 2019*
Accepted: *20 March 2020*
Published: *12 May 2020*

Citation:
Zhang J, Yang F, Jiang Y, Guo Y, Wang Y, Zhu X, Li J, Wan H, Wang Q, Deng Z, Xuan P and Yang W (2020) Preferential Subgenome Elimination and Chromosomal Structural Changes Occurring in Newly Formed Tetraploid Wheat—Aegilops ventricosa Amphiploid ($AABBD^vD^vN^vN^v$).
Front. Genet. 11:330.
doi: 10.3389/fgene.2020.00330

Artificial allopolyploids derived from the genera *Triticum* and *Aegilops* have been used as genetic resources for wheat improvement and are a classic example of evolution via allopolyploidization. In this study, we investigated chromosomes and subgenome transmission behavior in the newly formed allopolyploid of wheat group via multicolor Fluorescence *in situ* hybridization (mc-FISH), using pSc119.2, pTa535, and $(GAA)_7$ as probe combinations, to enabled us to precisely identify individual chromosomes in 381 S_3 and S_4 generations plants derived from reciprocal crosses between *Ae. ventricosa* ($D^vD^vN^vN^v$) and *T. turgidum* (AABB). A higher rate of aneuploidy, constituting 66.04–86.41% individuals, was observed in these two early generations. Of the four constituent subgenomes, D^v showed the highest frequency of elimination, followed by N^v and B, while A was the most stable. In addition, structural chromosomal changes occurred ubiquitously in the selfed progenies of allopolyploids. Among the constituent subgenomes, B showed the highest number of aberrations. In terms of chromosomal dynamics, there was no significant association between the chromosomal behavior model and the cytoplasm, with the exception of chromosomal loss in the D^v subgenome. The chromosome loss frequency in the D^v subgenome was significantly higher in the *T. turgidum* × *Ae. ventricosa* cross than in the *Ae. ventricosa* × *T. turgidum* cross. This result indicates that, although the D subgenome showed great instability, allopolyploids containing D subgenome could probably be maintained after a certain hybridization in which the D subgenome donor was used as the maternal parent at its onset stage. Our findings provide valuable information pertaining to the behavior patterns of subgenomes during allopolyploidization. Moreover, the allopolyploids developed here could be used as potential resources for the genetic improvement of wheat.

Keywords: tetraploid wheat, *Aegilops ventricosa*, amphiploid, chromosomal variation, mc-FISH

INTRODUCTION

Common wheat, or bread wheat, (*Triticum aestivum* L, AABBDD) is one of the most important food crops worldwide, with extensive reports on its speciation. To date, it has been widely accepted that two sequential allopolyploidization events occurred during evolution. The first allopolyploidization occurred about 0.3–0.5 million years ago, involving interspecific hybridization between *Triticum urartu* (AA) and an unknown *Aegilops* species, possibly related to *Aegilops speltoides* (SS), which led to the formation of emmer wheat (AABB) (Dvorak and Zhang, 1990; Huang et al., 2002; EI El Baidouri et al., 2017). Subsequently, allopolyploidization between tetraploid wheat (AABB) and *Aegilops tauschii* (DD) occurred about 10,000 years ago, which gave rise to the speciation of modern bread wheat (AABBDD) (Kilian et al., 2007; Marcussen et al., 2014). As a classical sample for the survey of evolution via allopolyploidization, several studies exist on chromosomal behavioral patterns in nascently synthesized allopolyploids of the *Triticum* tribe, including the crossing of tetraploid or hexaploid wheat with *Aegilops* and *Secale* species. For example, Zhao et al. (2011) observed that an individual plant harbored 50 chromosomes in the S_2 generation that were derived from a newly formed allohexaploid wheat line (2n = 42, AABBDD). Various frequencies of aneuploidy were detected in the S_1 and S_2 generations derived from the crosses between several genotypes of *Triticum durum* (AABB) and *Ae. tauschii* (DD) (Mestiri et al., 2010), and persistent aneuploidy was found to be associated with nascent allohexaploid wheat (AABBDD) in the S_1 to $>S_{20}$ generations (Zhang et al., 2013a). In addition to variation in chromosome numbers (chromosome loss/gain), extensive chromosome rearrangements, scale genomic changes of repetitive DNA, and copy-number variations in gene homologs were detected in four newly synthesized allotetraploid wheat lines (genome compositions were $S^{sh}S^{sh}A^mA^m$, S^lS^lAA, S^bS^bDD, and AADD, respectively) (Zhang et al., 2013b). Telosome mutations were also observed in a newly formed hexaploid wheat derived from a cross between *Triticum turgidum ssp. dicoccum* MY3478 (AABB) and *Ae. tauschii* SY41(DD) (Li H. et al., 2016). Among the constituent subgenomes, chromosomal instability exhibiting obvious subgenome-bias was observed. Most of the previous studies demonstrated a preferential elimination of the D subgenome compared to the A and B subgenome, and even the R, S, U, and Ns genome (Tiwari et al., 2010; Xie et al., 2012; Hao et al., 2013; Li et al., 2015; Guo et al., 2018). Mestiri et al. (2010) proposed that genome stability was dependent on the genotypes of both A, B, and D genome donors. Nonetheless, chromosomal alterations in the D subgenome, which are not directly contributed by wheat or *Ae. tauschii,* have not been investigated in detail, and it remains uncertain whether the chromosome behavior pattern exhibits cytoplasm-dependence.

Aegilops ventricosa Tausch (2n = 28, genome $D^vD^vN^vN^v$) has potential as a genetic source of germplasm due to its ability to tolerate biotic stresses and may be useful for wheat improvement. In recent decades, two sets of *Ae. ventricosa* introgression lines, H-93 and VPM1 derived from the interspecific hybridization of *T. aethiopicum* with *Ae. ventricosa* (accession AP-1) and *T. persicum* with *Ae. ventricosa* (accession #10), respectively (Maìa, 1967; Dosba et al., 1978), have been studied using cytogenetic methods and widely used in wheat improvement. Until this, H-93, VPM1, and their derivative lines have demonstrated resistances against several wheat pathogens, such as powdery mildew (Delibes et al., 1987), eyespot (*Pch1*) (Doussinault et al., 1983), rust (*Lr37*, *Yr17*, and *Sr38*) (Bariana and McIntosh, 1993, 1994; Bonhomme et al., 1995; Tanguy et al., 2005), cereal cyst nematode (*Cre2*, *Cre5,* and *Cre6*) (Delibes et al., 1993; Jahier et al., 2001; Ogbonnaya et al., 2001; Tanguy et al., 2005), and Hessian fly (*H27*) (Delibes et al., 1997). In order to effectively exploit the genetic value originating from *Ae. ventricosa*, novel *Ae. ventricosa* accessions are needed to be employed in the development of wheat-*Ae. ventricosa* introgression lines.

In this study, we used multicolor FISH (mc-FISH) to investigate 381 experimental plants, covering S_3 and S_4 generations derived from reciprocal crosses between *Ae. ventricosa* ($D^vD^vN^vN^v$) and *T. turgidum* (AABB). Our aim was to characterize the chromosome alterations in nascent wheat allopolyploids for use as novel genetic resources of *Ae. ventricosa*. As a result, chromosome alterations were found to be accompanied by the formation of the new allopolyploids. Subgenome and chromosome biases were also observed. No association was observed between chromosomal dynamics and cytoplasm, with the exception of the chromosomal loss in the D^v subgenome.

MATERIALS AND METHODS

Plant Materials

Aegilops ventricosa cv. RM271 (D subgenome of *Ae. ventricosa,* or D^v) was supplied by Prof. Lihui Li (Institute of Crop Sciences, Chinese Academy of Agricultural Sciences). *Aegilops tauschii* SQ 665 (D subgenome of *Ae. tauschii*, or D^t) was provided by International Maize and Wheat Improvement Center (CIMMYT). *T. turgidum* var. *durum* cv. Langdon and common wheat Chinese Spring (CS) (D subgenome of bread wheat, or D^b) were conserved by our library. These four experimental lines were used to develop a standard FISH karyotype of A, B, D^v, D^t, D^b, and N^v subgenomes. Spontaneous amphidiploids of *T. turgidum–Ae. ventricosa* were obtained from fertile F_1 generations of *Ae. ventricosa* cv. RM271 (as female or male) crossed with *T. turgidum* cv. Langdon (*T. turgidum* × *Ae. ventricosa*; *Ae. ventricosa* × *T. turgidum*) via chromosome autoduplication. The seeds of S_3 [derived from *T. turgidum* × *Ae. ventricosa* (group 1, or G_1) and *Ae. ventricosa* × *T. turgidum* (group 2, or G_2)] and S_4 [derived from *T. turgidum* × *Ae. ventricosa* (group 3, or G_3) and *Ae. ventricosa* × *T. turgidum* (group 4, or G_4)] generations were collected. Then, 124, 53, 103, and 101 seeds (381 seeds in total) were randomly selected from G_1, G_2, G_3, and G_4, respectively, for analysis of the FISH karyotype (**Figure 1**).

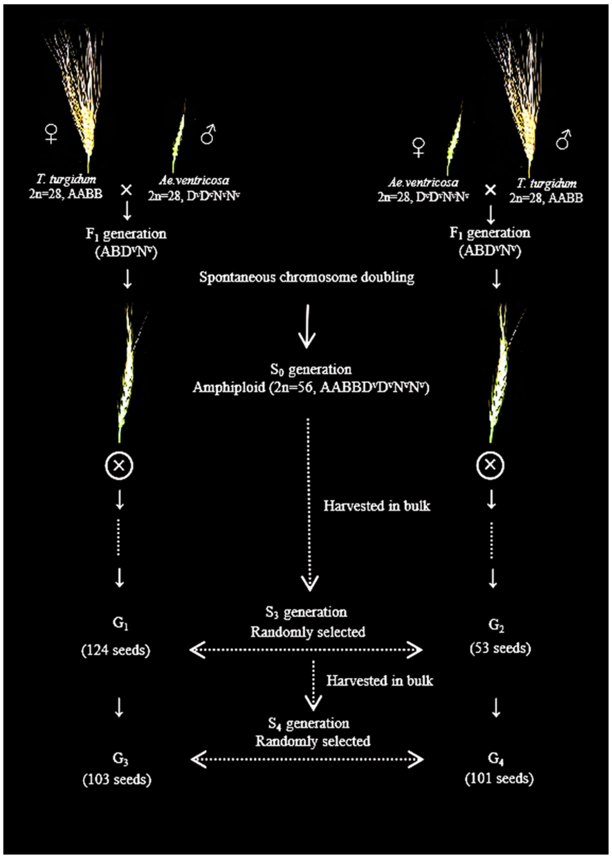


FIGURE 1 | Schematic of strategy used to obtain the experimental seeds (G_1, G_2, G_3, and G_4) using hybridization between *T. turgidum* and *Ae. ventricosa.*

mc-FISH Procedures

The root tips used for karyotype analyses were collected from germinating seeds and treated with nitrous oxide. The protocol was conducted as described by Zhang J. et al. (2018) The chromosomal preparation of mitotic metaphases was performed as reported by Kato et al. (2004). To identify every pair of chromosomes originating from *T. turgidum* and *Ae. ventricosa*, oligo-nucleotide probes Oligo-pSc119.2, Oligo-pTa535, and Oligo-$(GAA)_7$ were labeled on the 5′-end with 6-carboxyfluorescein (6-FAM), 6-carboxytetramethylrhodamine (Tamra), and Cy5, respectively, and used as described by Tang et al. (2014) and Li G. et al. (2016) (Invitrogen, Shanghai, China). FISH protocols were adapted from the methods described by Han et al. (2006). The slides were stored in a moist box at 37°C for 4 h and washed in 2 × SSC at room temperature. The slides were mounted with VECTASHIELD mounting medium with 4′,6-diamidino-2-phenylindole (DAPI) (Vector Laboratories, CA, United States). Images were captured using a Leica DM2500 fluorescence microscope (Leica, Wetzlar, Germany) equipped with a cooled charge-coupled device camera (Leica, Wetzlar, Germany) operated with LAS Live software (version 4.6) (Leica, Wetzlar, Germany). Signal pattern images for the same metaphase cell were taken over two rounds. The first round was used to capture the signals of probes labeled with 6-FAM (Oligo-pSc119.2) and Tamra (Oligo-pTa535), which were displayed green and red and the chromosomal background stain, DAPI, which was displayed as blue; the three resulting images (Oligo-pSc119.2, Oligo-pTa535 and DAPI) were merged. In the second round, the signals of Oligo-$(GAA)_7$ labeled with Cy5 and the chromosomal background stain, DAPI, which displayed yellow and red, respectively, were obtained; the two resulting images (Oligo-$(GAA)_7$ and Cy5) were merged.

Statistics Analysis

Statistical analyses were performed using SPSS version 19.0 (IBM, United States) and GraphPad Prism version 8.01 (GraphPad Software, United States). Chi-square test or Student's *t*-test were used to estimate the statistical significance for each comparison with a *p*-value 0.05 as the threshold.

RESULTS

Amphiploid Plant Morphology (S_1 Generation)

Phenotypically, amphiploids exhibited a spike-shape similar to that of their parents *Ae. ventricosa.* The spike-lengths of amphiploids were clearly longer than those either of the parents, *T. turgidum* and *Ae. ventricosa.* The short awns of the spikes in the progenies were likely inherited from *Ae. ventricosa* (**Figure 2A**). In addition to the spikes, the length and width of the amphiploids seeds were higher than those of either of the parents. Moreover, the amphiploids had tough glumes similar to those of the parent *Ae. ventricosa*, which made the spikes difficult to thresh (**Figure 2B**).

FISH Karyotypes of A, B, D^v, N^v Subgenomes Were Established Using the Probe Combination of Oligo-pSc119.2, Oligo-pTa535, and Oligo-$(GAA)_7$

The chromosomes of the parents *T. turgidum* and *Ae. ventricosa* were analyzed by mc-FISH using a combination of probes Oligo-pSc119.2, Oligo-pTa535, and Oligo-$(GAA)_7$. As shown in **Figures 3A–D**, I, Oligo-pSc119.2 was mainly localized to the B and N^v genomes, while high levels of Oligo-pTa535 signal were observed on A and D^v genomes. Oligo-$(GAA)_7$ exhibited strong signals in the centromeric regions of all the B chromosomes and large parts of the D^v and N^v chromosomes. The integrated use of probes allowed for the identification of the individual chromosomes of the A, B, D^v, and N^v genomes. *Ae. ventricosa* is a tetraploid originating from of *Ae. tauschii* and *Ae. uniaristata* (NN). The D^v genome of *Ae. ventricosa* ($D^vD^vN^vN^v$) is similar to that of *Ae. tauschii* (DD) (McNeil et al., 1994), which allowed us to precisely identify each pair of D^v and N^v chromosomes according to the FISH karyotype of *Ae. tauschii* and the descriptions reported by Badaeva et al. (2011). Comparison of the FISH karyotypes of the D^v genome with that of D^t genome revealed significant differences in terms of the microsatellite sequence

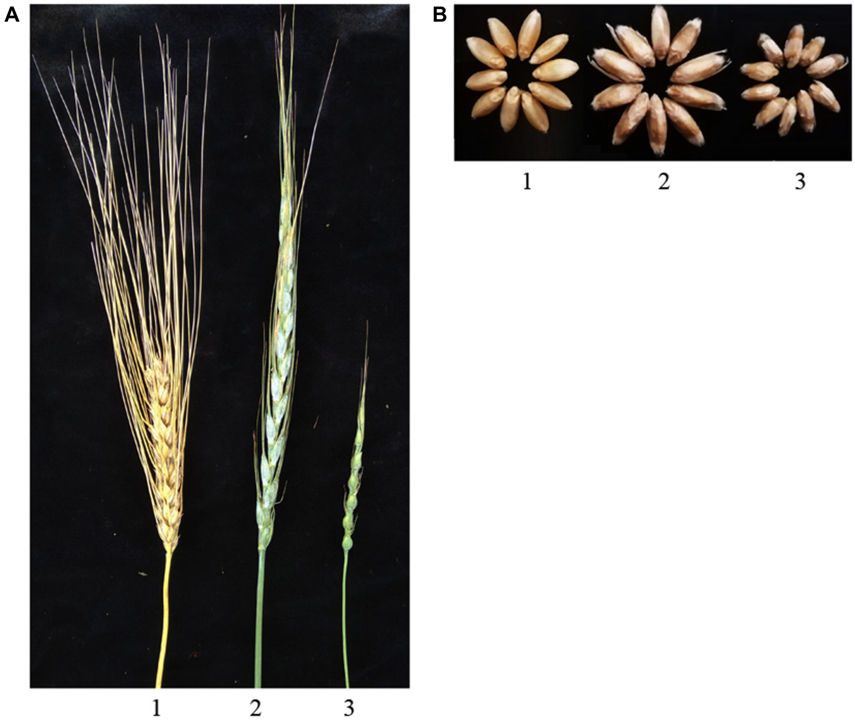


FIGURE 2 | Spike **(A)** and seeds **(B)** morphology. 1–3 are *T. turgidum* var. *turgidum* cv. Langdon, amphiploid, and *Ae. ventricosa* cv. RM271, respectively.

distribution patterns, based on the signal distribution of Oligo-$(GAA)_7$. For example, $1D^v$ harbored strong and weak signal bands at the end of long and short arms, respectively, while $1D^t$ showed weak signal bands at both ends of the long and short arms; $3D^v$ carried signals at the centromeric region and the end of short arm, while $3D^t$ carried signals at the centromeric region and the end of the long arm; $4D^v$ showed a stronger signal than $4D^t$; $6D^v$ had an obvious signal at the centromeric region and sub-terminal region of the long arm. Conversely, $6D^t$ showed the extremely low levels of signal along the chromosome; $7D^v$ showed signal at the centromeric region, whereas $7D^t$ showed signal at the terminal region of the chromosome. Compared to the D^v and D^t genomes, extremely low levels of signal of Oligo-$(GAA)_7$ was detected along the all chromosomes from the D^b genome of bread wheat, Chinese Spring (CS) (**Figures 3C–I**). These results indicated that the signal patterns of Oligo-$(GAA)_7$ of D^v and D^t were similar to those of D^b (**Figure 3I**).

High Levels of Whole-Chromosome Aneuploidy in Early Generations

The chromosome composition of the 381 individuals from the four groups were analyzed via mc-FISH. All of the mitotic cells investigated showed the same FISH karyotype, indicating that no somatic alteration existed among the cells from the same individuals. A small proportion of euploids were identified (**Figure 4**), and high levels of whole-chromosome aneuploidy were observed in all groups, with frequencies varying from 67.74 to 86.41% (**Figure 5A**). The frequency of aneuploids was significantly higher than that in euploids (t-test, p = 0.000) (**Figure 5B**). All four groups showed variable chromosome numbers and the chromosome numbers of the experimental plants ranged from 46 to 57 (**Figure 6** and **Supplementary Figure S1**). Of the 381 plants, the frequency of plants exhibiting chromosomal loss was prominent compared to those exhibiting chromosomal gain, with rates of 72.23 and 4.46%, respectively.

We observed a special type of aneuploidy in G_2, and G_3. This type of aneuploidy contained 56 chromosomes, as in euploids, but exhibited chromosome loss and gain, which was denoted as hidden aneuploidy by Zhang et al. (2013a). Of the 381 plants, two plants (0.52%) were classified as having "hidden aneuploidy," 17Y-44-95 (in G_2) and Y1701-1-2 (in G_3) (FISH signal patterns are shown in **Figures 7** A–D). Y1701-1-2 involved 7N loss/1D gain, while 17Y-44-95 showed a complex "hidden aneuploidy" pattern involving 1A, 6A, 7N loss/7A, 3A, 1B gain. Notably, the two "hidden aneuploidy" lines involved chromosomes (loss/gain) originating from different homologous groups.

These results indicate a large incidence of aneuploidy in the early generations of the neoallopolyploids resulting from the *T. turgidum* × *Ae. ventricosa* cross. Subsequently, we explored whether chromosome loss/gain events were associated with genome biases. D^v subgenome was found to show the highest frequency of chromosomal loss (50.27%) (Chi-square test, $p = 0.00$). The N^v subgenome (30.59%) showed a significantly higher frequency of chromosomal loss than the B (13.81%) or A (11.61%) subgenomes (Chi-square test, p = 0.017, N^v vs. B; Chi-square test, p = 0.02, N^v vs. A), while no significant differences were detected between the A and B subgenomes (Chi-square test

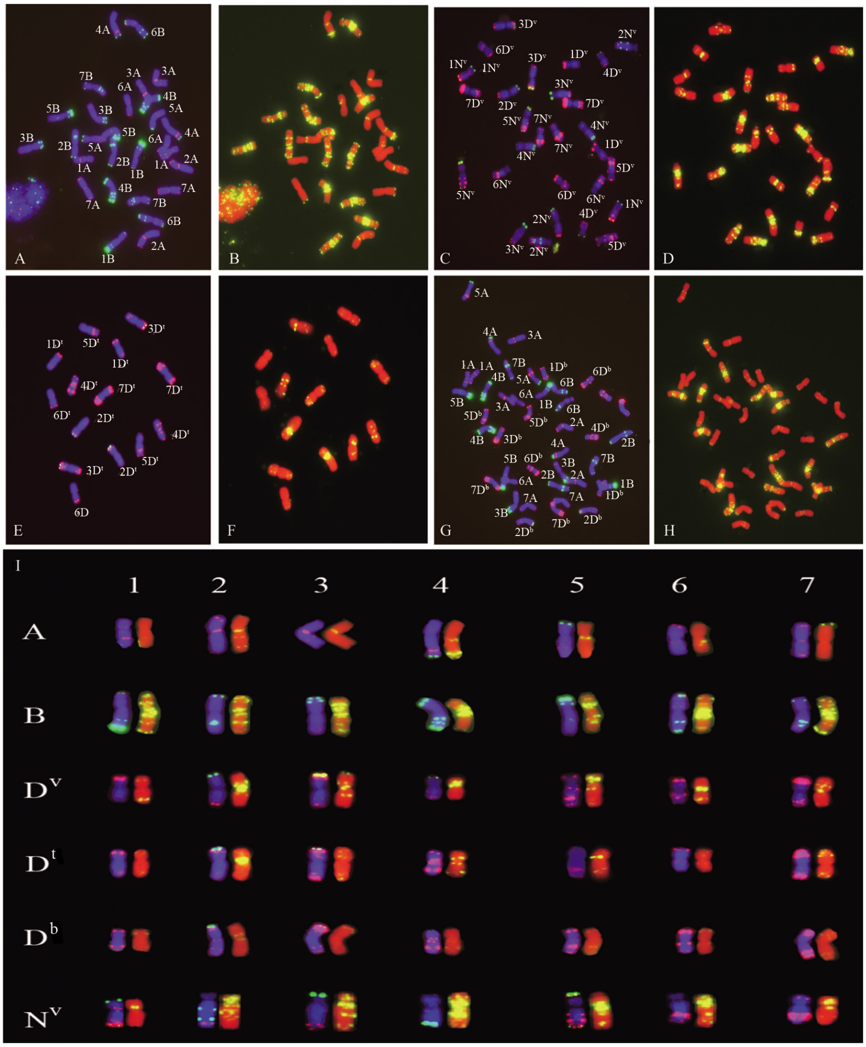


FIGURE 3 | FISH analysis of *T. turgidum* cv. Langdon **(A,B)**, *Ae. ventricosa* cv. RM271 **(C,D)**, *Ae.tauschii* SQ 665 **(E,F)**, common wheat Chinese Spring (CS) **(G,H)**, and the karyograms of the A, B, D^{v}, D^{t}, D^{b}, and N^{v} subgenomes **(I)**. **(A,C,E,G)** chromosomes stained by DAPI (blue), oligo-nucleotides Oligo-pSc119.2 (green), and Oligo-pTa535 (red). **(B,D,F,H)** Chromosomes staining with DAPI (red) and Oligo-(GAA)$_7$ (yellow). A, B in panel **(I)** are the subgenomes of *T. turgidum* cv. Langdon; D^{v} and N^{v} are the subgenomes from *Ae. ventricosa* cv. RM271; D^{t} is the genome originated from *Aegilops tauschii* cv. SQ 665; D^{b} is the subgenome from bread wheat Chinese Spring (CS).

p = 0.415) (**Figure 8**). In terms of chromosome gain, A and D^{V} showed the highest frequency, 3.48 and 3.29%, respectively, while N^{V} showed the lowest frequency of chromosomal gain (0.99%). However, no significant differences related to chromosomal gains were found among the four subgenomes (Chi-square test, $p > 0.05$) (**Figure 8**).

Incidence of Chromosome Structural Variations and Different Propensities Among Constituent Genomes

Compared to variations in chromosome numbers, variations in chromosome structures occurred in a smaller proportion. Among the 381 plants, changes in chromosome structures were observed in 39 individuals, with a rate of 10.24%, including 47 events of chromosomal breakage. Among these 39 plants, two types of chromosomal structural variation were detected: chromosomal breakage (including 1AS. 1AL-, 3AL, -4AS.4AL, 6AL, 1BL, 2BS. 2BL-, 4BL, 5BS, -6BS.6BL, 7BS. 7BL-, 2D^{V}L, 3D^{V}L, 6D^{V}S.6D^{V} L-, 1N^{V}L, 3N^{V}S.3N^{V} L-, and 6N^{V}S) and translocation (including 3BL.3BL, 5N^{V}L.5N^{V}L, 7A.7D, 5AL.5AL, and 7N^{V}L.3BL).

All four subgenomes (A, B, D^{V}, and N^{V}) underwent chromosome structural variation, and 11 (2.89%, A subgenome), 23 (6.04%, B subgenome), five (1.31%, D^{V} subgenome), and eight (2.10%, N^{V} subgenome) plants carrying chromosomal structural change, respectively, were observed in 381 plants (**Figure 9A**). Similar to the variation in chromosome numbers, chromosomal structural variations were accompanied by genome bias. Among the four constituent subgenomes, B subgenome showed significantly higher frequency of structural variations

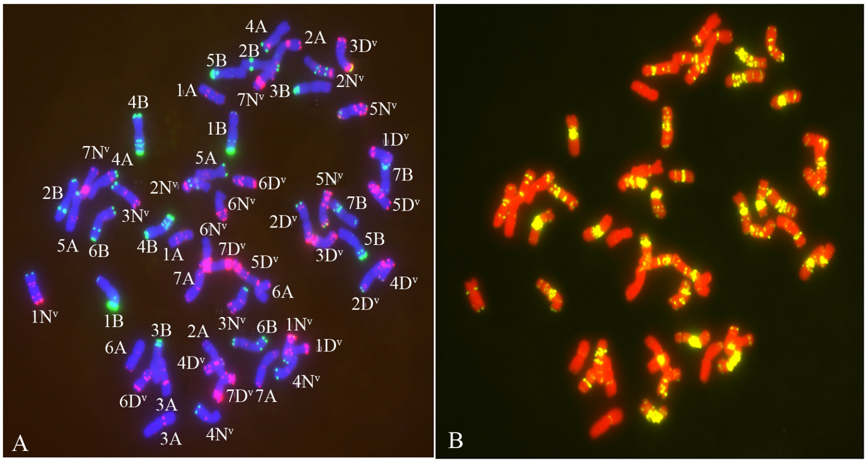


FIGURE 4 | The FISH pattern from euploids. **(A)** Chromosomes staining with DAPI (blue) oligo-nucleotides Oligo-pSc119.2 (green), Oligo-pTa535 (red). **(B)** chromosomes staining by DAPI (red) and Oligo-$(GAA)_7$ (yellow).

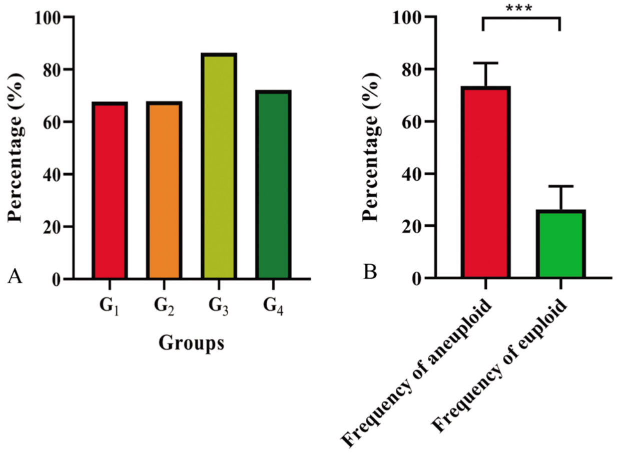


FIGURE 5 | Aneuploid distribution **(A)** in four groups and frequencies of aneuploids and euploids **(B)**. *P*-values labeled with asterisks denote significant difference at $^{***}P < 0.001$.

(Chi-square test, $p = 0.048$, B vs. A; $p = 0.005$, B vs. D^v; $p = 0.016$, B vs. N^v). A large proportion of chromosomes were observed to exhibit structural variations from all four subgenomes, while no visible structural variations were detected on chromosomes 2A, $1D^v$, $4D^v$, $5D^v$, $2N^v$, and $4N^v$ (**Figure 9B** and **Supplementary Figure S2**).

Significant Differences in Chromosomal Behavior Model Between Progenies From Reciprocal Crosses Observed Only in the Chromosome Loss of the D^v Subgenome

In terms of the aneuploidy frequency and subgenome bias of chromosomal numerical/structural variation, no significant differences were found between the tetraploid wheat cytoplasm and *Ae. ventricosa* cytoplasm (**Figures 10A–F,H**). Interestingly, a significant difference was detected for chromosome loss in D^v subgenome between the reciprocal crosses (*t*-test, $p = 0.006$ for *T. turgidum* × *Ae. ventricosa* vs. *Ae. ventricosa* × *T. turgidum*) (**Figure 10G**). However, because the frequencies were low, we were unable to conduct a statistical analysis for chromosome gain and chromosomal structural variations of the four constituent subgenomes.

DISCUSSION

Repetitive sequences have significantly contributed to our understanding of genome organization and recombination during the evolution of plants (Friesen et al., 2001). Numerous FISH probes have been developed based on repetitive sequences (Cuadrado and Jouve, 2010; Tang et al., 2014, 2016, 2018; Fu et al., 2015; Lang et al., 2018; Liu et al., 2018), which are widely used for the identification of chromosomes of the *Triticum* genus (Badaeva et al., 2011; Zhang et al., 2013b; Delgado et al., 2016; Zhao et al., 2018; Ren et al., 2019). In this study, we established the FISH karyotypes of the A, B, D^v, and N^v subgenomes using Oligo-pSc119.2, Oligo-pTa535 and Oligo-$(GAA)_7$. As expected, these probes distinguished between the A, B, D^v, and N^v subgenomes. Moreover, the signal patterns on the chromosomes of the A and B subgenomes were consistent with the results reported by Tang et al. (2014). Since *Ae. ventricosa* originated from the hybridization of *Ae. tauschii* (DD) and *Ae. uniaristata* (NN) (McNeil et al., 1994), each individual chromosome from D^v subgenome could be identified according to the FISH signal patterns of D genome of *Ae. tauschii*.

In the present study, we compared the FISH karyotypes of the three types of D genomes (D^t genome of *Ae. tauschii*, D^v of *Ae. ventricosa*, and D^b of *T. aestivum*), and we found that the three D genomes showed slight differences within themselves. These results indicated that D genome present in

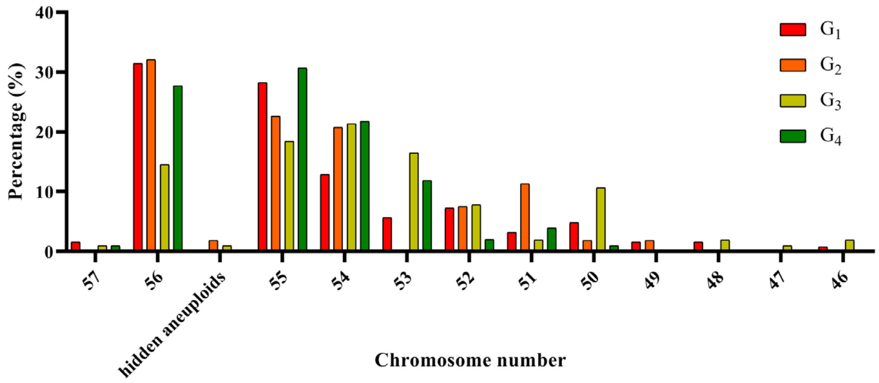


FIGURE 6 | Distribution of the number of chromosomes in the four experimental groups.

Ae. ventricosa (D^v) and *T. aestivum* (D^b) underwent small modifications compared to the ancestral D genomes of *Ae. tauschii*, which is in agreement with the results reported by Mirzaghaderi and Mason (2017). Notably, we found that the three D genomes showed similar FISH signal patterns for the Oligo-pSc119.2 and Oligo-pTa535 probes compared to the Oligo-$(GAA)_7$ probe, which indicated that the tandem repeat sequences of pSc119.2 (KF719093) and pTa535 (KC290894.1) were likely to be more conserved than simple repeat sequences $(GAA)_n$ over evolutionary time. Interestingly, the $1D^v$–$7D^v$, $2D^t$–$5D^t$, and $7D^t$ chromosomes showed observable levels of Oligo-$(GAA)_7$ signal patterns. Specifically, $2D^v$ and $2D^t$, $4D^v$ and $4D^t$, and $5D^v$ and $5D^t$ shared similar Oligo-$(GAA)_7$ signal patterns. However, the chromosomes of D^b subgenome showed fewer and lower levels of Oligo-$(GAA)_7$ signals compared to the chromosomes of the D^v and D^t subgenomes. The FISH signal patterns of Oligo-$(GAA)_7$ on the chromosomes of the D^v and D^t subgenomes were more similar than those of the chromosomes of the D^b and D^t subgenomes. These results showed a burst of simple repeat sequences occurred on both D^v and D^t subgenomes rather than D^b subgenome, implying that the divergence between the D^t of *Ae. tauschii* and D^v of *Ae. ventricosa* most likely occurred later than between D^t of *Ae. tauschii* and D^b of bread wheat, indicating that intraspecific hybridization of *Ae. tauschii* and *Ae. uniaristata* (NN) may have occurred later than the intergeneric hybridization of *Ae. tauschii* and emmer wheat.

Studies on the behavior of subgenomes and single chromosomes concerning the last post-allopolyploidization timescale are valuable to improve our understanding of the formation and evolution of allopolyploids. Recently, many studies have demonstrated that allopolyploidization in the wheat group triggered extensive genomic shocks, which led to karyotype changes, including changes in chromosome structure and numbers (Hao et al., 2013; Zhang et al., 2013a; Li G. et al., 2016; Guo et al., 2018). Large and frequent whole chromosome loss/gain events and chromosome/subgenome biases resulting from aneuploidy in the neoallopolyploids of the wheat group have been described in detail in the literature (Tiwari et al., 2010; Zhou et al., 2012; Hao et al., 2013; Zhang et al., 2013b; Li et al., 2015). However, the majority of the literature is only concerned with neoallopolyploids harboring the D subgenome originating from *Ae. tauschii* or *T. aestivum.* As a result, neoallopolyploids carrying other types of D subgenome originating from *Aegilops* genus have not been fully investigated; whether the maternal parent effect the chromosomal behavior remains unknown.

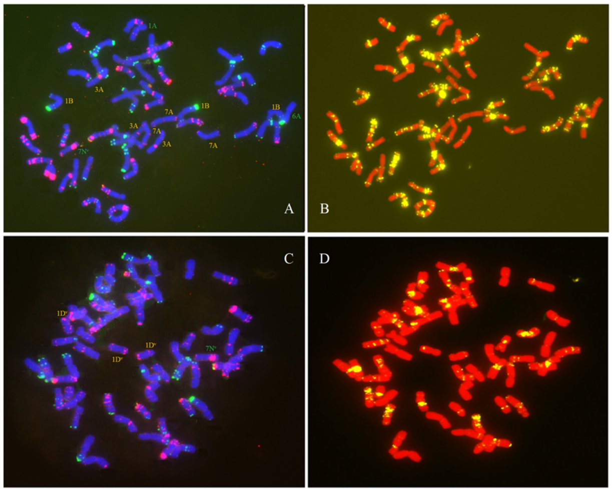


FIGURE 7 | FISH analysis of two types of "hidden aneuploidy" using Oligo-pSc119.2 (green), Oligo-pTa535 (red) and Oligo-$(GAA)_7$ (yellow) as probes. **(A,B)** 17Y-44-95; **(C,D)** Y1701-1-2. Green denotes chromosome loss and yellow denotes chromosome gain.

We addressed these issues using two sets of newly formed *T. turgidum-Ae. ventricosa* allopolyploidy ($AABBD^vD^vN^vN^v$) lines derived from reciprocal crosses between *Ae. ventricosa* and *T. turgidum.* Many aneuploids were observed in the S_3 (G_1 and G_2) and S_4 (G_3 and G_4) generations, with frequencies of 67.74, 67.92, 86.41, and 72.28%, respectively, which correlate with the results of several recent studies (Zhang et al., 2013b; Guo et al., 2018). It is widely believed that aneuploidy generally occurs in newly formed allopolyploid wheat lines. However, natural wheat varieties show a much lower frequency of aneuploid individuals (1–3%) (Riley and Kimber, 1961). Recently, several studies have attempted to address the underlying mechanisms of stability that

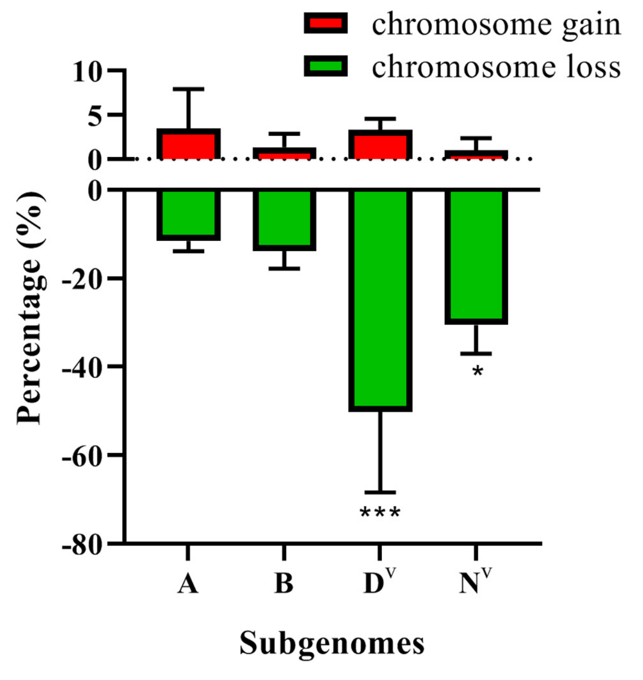


FIGURE 8 | The frequencies of subgenome loss and gain. *P*-values labeled with asterisks denote significant differences at $^{*}P < 0.05$ and $^{***}P < 0.001$.

give rise to the formation of allopolyploid in wheat. For example, Zhang et al. (2013b) found that persistent aneuploidy is generally associated with nascent allohexaploid wheat over multiple generations (S_1–S_{20}) and further proposed that karyotypic stabilization could not be achieved even by consecutive screening for euploidy. Mestiri et al. (2010) demonstrated that the stability of synthetic allohexaploids in wheat was dependent on the variability of genotypes of *T. turgidum* and *Ae. tauschii.* Our findings showed that the cytoplasm of the maternal parent did not affect the frequency of euploidy in the wheat group, which suggests that many aneuploids would occur in early generations no matter which plant was used as the maternal parent. Thus, the key constituent affecting the stabilization of allopolyploids in natural wheat also warrants further investigation. Given the much higher rate of the plants harboring chromosome loss (72.23%) compared to chromosome gain (4.46%), it was obvious that chromosome loss would occur more easily than chromosome gain, which most likely depends on the capacity of cell nucleus *per se*.

Chromosomal aneuploidy is known to exhibit preferential chromosome elimination. Our results demonstrated that the D^v subgenome showed a higher frequency of loss (50.27%) and gain (3.29%) than the other three constituent subgenomes, suggesting that D^v is the most unstable subgenome of our newly formed allopolyploids. The result is in agreement with a significant number of previous studies, such as wheat-*Secale* (RR) allopolyploids (A, B, D, and R) (Dou et al., 2006; Zhou et al., 2012; Hao et al., 2013; Li et al., 2015), wheat-*Ae.kotschyi* ($U^kU^kS^kS^k$) allopolyploids (A, B, D, U^k, and S^k) (Tiwari et al., 2010), newly formed wheat allotetraploids (AADD) (Guo et al., 2018), and trigeneric hybrids (A, B, D, R, and Ns) of wheat–*Secale* (RR)–*Psathyrostachys huashanica* (NsNs) (Xie et al., 2012), but contrary to the "pivotal-differential genome evolution hypothesis"(Mirzaghaderi and Mason, 2017), which proposed that D, A, and U were pivotal genomes, and underwent little modifications. Thus, they were considered to be more stable than under-dominant subgenomes (the other subgenomes harbored in the wheat group with the exception of A, D, and U). Possibly, this mechanism may have given rise to the formation of allohexaploid wheat over evolutionary time (Zhang et al., 2013a). There is significant evidence that has supported pivotal-differential genome patterns as a common phenomenon during the evolutionary process of wheat group via allopolyploidization. For example, allotetraploids *Ae. crassa* (DM), *Ae. cylindrica* (DC), and *Ae. triuncialis* (UC) showed lower amounts of sequence loss in their pivotal subgenomes (D and U) than in their differential subgenomes (M and C) (Senerchia et al., 2014); in terms of allohexaploid bread wheat, the higher sequence order conservation was detected in the A genome relative to B and in the D genome compared to A and B (Pont et al., 2013); in allohexaploid *Ae. neglecta* ($2n = 6 \times = 42$, $U^nU^nM^nM^nN^nN^n$), U (pivotal subgenome) has remained mostly intact relative to M and N (differential subgenome) (Badaeva et al., 2004). Integrating our findings with the previous results, we propose that in early generations of newly resynthesized allopolyploids of wheat group, chromosomal behavior pattern probably may not strictly follow this rule. The D subgenome is likely to be more easily "shocked" than other subgenomes (including pivotal subgenomes, A, B, and U, even the differential subgenomes S, R, N, and Ns) once they are merged with other genomes, which leads to its preferential elimination. The preferential elimination or stability of D subgenome in the genetic background of the newly formed allopolyploids may depends on its genotypes (Mestiri et al., 2010), which accounts for why the D subgenome exhibited preferential elimination, while in another research, the D subgenome showed the highest stability as reported by Zhang et al. (2013a). Additionally, we found that the frequency of chromosome loss in the D^v genome from the *T. turgidum* × *Ae. ventricosa* cross was significantly higher than that observed in the *Ae. ventricosa* × *T. turgidum* cross. This may be due to the fact that the cytoplasm of the maternal parent gas different effects on different subgenomes. This finding suggests that although the D subgenome showed the greatest instability, allopolyploids containing D subgenome are likely to be maintained when the D subgenome donor was used as the maternal parent at its onset stage.

Small proportions of "hidden aneuploids" (0.78%) exhibiting the expected chromosomal number of parental euploids ($2n = 56$) with no discernible chromosomal structural alterations, but with loss and gain of chromosomes, were found. In the current study, none of the "hidden aneuploids" were found to involve homologous chromosomes, and the events of chromosomal loss/gain likely occurred randomly, which was consistent with the results described by Zhang et al. (2013a). The possible reason for this phenomenon is the allopolyploid nature of the wheat group.

We also observed large chromosomal structural changes, which is in line with the results reported by several studies

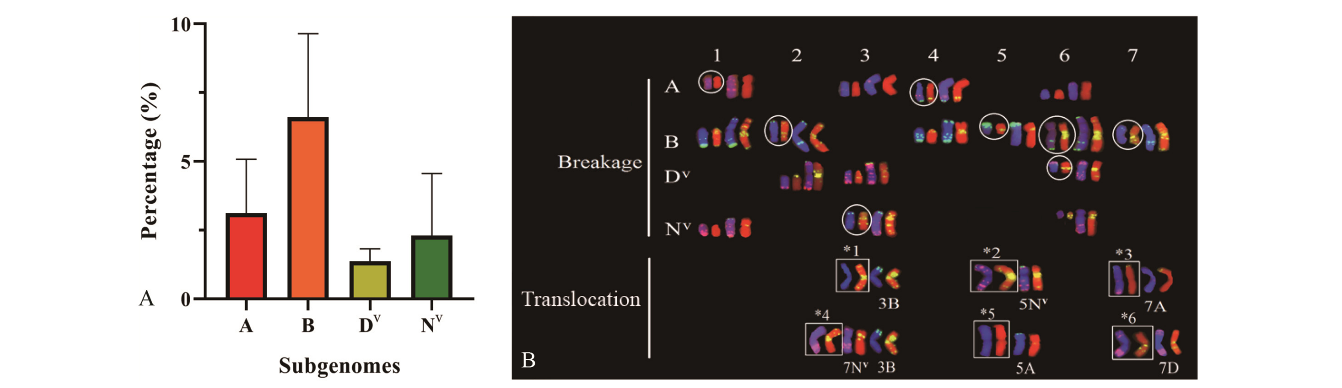


FIGURE 9 | The distribution of chromosomal structural variations of four subgenome **(A)**, and the FISH karyotype of chromosomal structural changes **(B)** by use of chromosomal structural variations. *1, 3BL.3BL; *2, 5N^{V}L.5N^{V}L; *3, 7AS.7D^{V}L; *4, 7N^{V}L.3BL; *5, 5AL.5AL; *6, 7D^{V}S.7AL. Circles denote chromosomes with breakages occurring in non-centromeric region.

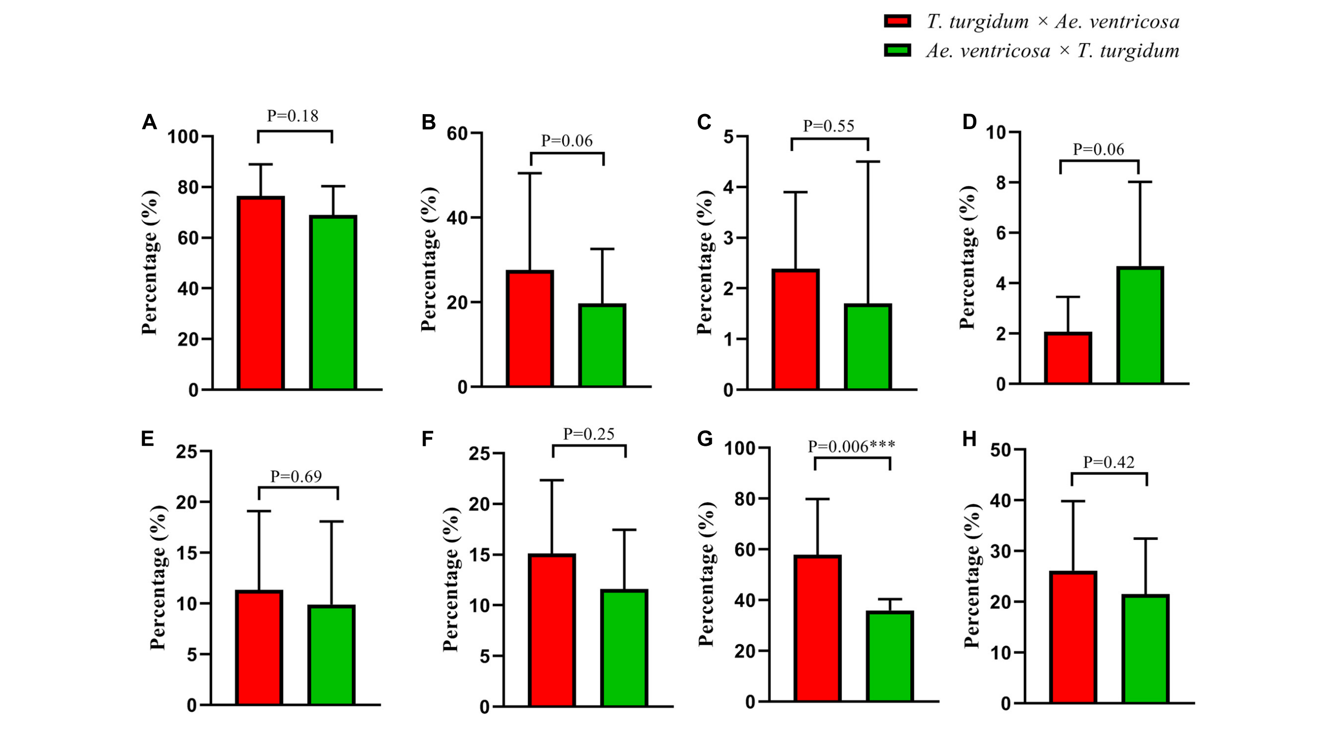


FIGURE 10 | Chromosomal behavior biases in reciprocal crosses. **(A)** Frequencies of aneuploids; **(B)** Frequencies of chromosome loss; **(C)** Frequencies of chromosome gain; **(D)** Frequencies of chromosomal structural variation; **(E–H)** Frequencies of chromosome loss in A, B, D^{V}, and N^{V} subgenome.

(Hao et al., 2013; Fu et al., 2015; Su et al., 2015; Li G. et al., 2016; Guo et al., 2018). In contrast, few or no chromosomal structural changes have been identified in other studies (Mestiri et al., 2010; Zhao et al., 2011; Zhang et al., 2013a). In terms of the possible reasons for this discrepancy, it is possible that there was incompatibility between the subgenomes of the wheat group and the alien genomes (such as R genome), which would have led to the disorganization of meiotic pairing and given rise to chromosomal structural alterations. In addition, our findings revealed that the B subgenome showed the highest frequencies of chromosomal structural variation (6.04%), which was likely due to its high heterochromatin content and large genome size (Zhang et al., 2015).

In this study, we characterized the chromosomal behavior of early generations of the newly formed *T. turgidum-Ae. ventricosa* allopolyploids. Plants carrying numerical and structural chromosomal variations were found, indicating that genetic variations may occur on the post-allopolyploidization timescale. These variations are likely to confer rich adaptive abilities to individuals in various natural habitats, thus, fueling the establishment of new species (Han et al., 2015). As such, our results provide insight into the possible reasons for the existence of aneuploidy in the early generations of allopolyploidization in wheat group. Chromosome loss in the D^v subgenome showed cytoplasm-dependence, indicating a possible mechanism of allopolyploids harboring D subgenome. The synthetic amphiploids of the wheat group could also be used as a "bridge" for the transfer of valuable alien genes to wheat (Zhang et al., 2017; Zhang D. L. et al., 2018). In this respect, our *T. turgidum-Ae. ventricosa* amphiploids, possessing appealingly large seeds, could be used as a potential genetic resource for wheat improvement.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

WY carried out the conceptualizaton. JZ and QW carried out the data curation. JZ, FY, YG, YW, and XZ performed the formal analysis. JZ, YG, and WY carried out the funding acquisition. JZ, YJ, and JL carried out the investigation. YW carried out the methodology. PX carried out the project administration. ZD performed the validation. JZ wrote the original draft. HW wrote, reviewed, and edited the manuscript.

FUNDING

We thank the National Nature Science Foundation of China (No. 31801362), Applied Basic Research Programs of Science and Technology Department, Sichuan Province (Key Project) (No. 2019YJ0605), the Sichuan Wheat Breeding Community (No. 2016NYZ0030), Project of Innovation Ability Improvement, Sichuan Province (No. 2019KXJJ-001), Foundation for Excellent Thesis of Sichuan Academy of Agricultural Sciences (No. 2018LWJJ-017), Applied Basic Research Programs of Science and Technology Department, Sichuan Province (Key Project) (No. 2018JY0629) and Crop Molecular Breeding Platform of Sichuan Province (No. 2016NZ0103) for financial support.

ACKNOWLEDGMENTS

We thank Dr. Zujun Yang, School of Life Sciences and Technology, University of Electronic Science and Technology of China, for careful reviewing and helpful comments, and we also thank Mr. Dalian Xu, Chengdu Time Measuring Instruments Corporation Ltd., for providing us nitrous oxide gas chamber.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00330/full#supplementary-material

FIGURE S1 | FISH analysis for metaphase spreads with chromosomal numerical variations. **(A–L)** Metaphase spreads with 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, and 57 chromosomes, respectively. 1s: Metaphase spread investigated using Oligo-pSc119.2 (green) and Oligo-pTa535 (red) as probes, 2s: Metaphase spreads using Oligo- (GAA)$_7$ (yellow) as probe. Chromosomes were stained with DAPI (blue on 1s and red on 2s). Green denotes chromosome loss and yellow denotes showed chromosome gain. The two chromosomes lost are listed at the lower right corner of every images.

FIGURE S2 | FISH analysis for metaphase spreads with chromosomal structural variations. **(A–T)** Metaphase spreads with 1AS. 1AL-, 3AL, -4AS.4AL, 7BS. 7BL-, 6AL, 1BL, 2BS. 2BL-, 4BL, 5BS, -6BS.6BL, 2D^{v}L, 3D^{v}L, 6D^{v}S.6D^v L-, 1N^{v}L, 3N^{v}S.3N^v L-, 6N^{v}S, 3BL.3BL, 5N^{v}L.5N^{v}L, 7A.7D^v, 5AL.5AL, and 7N^{v}L.3BL, respectively. 1s: Metaphase spread investigated using Oligo-pSc119.2 (green) and Oligo-pTa535 (red). 2s: Metaphase spreads investigated using Oligo- (GAA)$_7$ (yellow). Chromosomes were stained with DAPI (blue on 1s and red on 2s). Red denotes type of chromosomal structural variations.

REFERENCES

Badaeva, E. D., Amosova, A. V., Samatadze, T. E., Zoshchuk, S. A., Shostak, N. G., Chikida, N. N., et al. (2004). Genome differentiation in *Aegilops*. 4. Evolution of the U-genome cluster. *Plant Syst. Evol.* 246, 45–76.

Badaeva, E. D., Dedkova, O. S., Zoshchuk, S. A., Amosova, A. V., Reader, S. M., Bernard, M., et al. (2011). Comparative analysis of the N-genome in diploid and polyploid *Aegilops* species. *Chromosome Res.* 19, 541–548. doi: 10.1007/s10577-011-9211-x

Bariana, H. S., and McIntosh, R. A. (1993). Cytogenetic studies in wheat. XV. Location of rust resistance genes in VPM1 and its genetic linkage with other disease resistance genes in chromosome 2A. *Genome* 36, 476–482. doi: 10.1139/g93-0659

Bariana, H. S., and McIntosh, R. A. (1994). Characterization and origin of rust resistance and powdery mildew resistance in VPM1. *Euphytica* 76, 53–61. doi: 10.1007/BF00024020

Bonhomme, A., Gale, M. D., Koebner, R. M. D., Nicolas, P., Jahier, J., and Bernard, M. (1995). RFLP analysis of an *Aegilops ventricosa* chromosome that carries a gene conferring resistance to leaf rust (*Puccinia recondita*) when transferred to hexaploid wheat. *Theor Appl Genet.* 90, 1042–1048. doi: 10.1007/BF00222919

Cuadrado, Á., and Jouve, N. (2010). Chromosomal detection of simple sequence repeats(SSRs) using nondenaturing FISH (ND-FISH). *Chromosoma* 119, 495–503. doi: 10.1007/s00412-010-0273-x

Delgado, A., Carvalho, A., Martín, A. C., Martín, A., and Lima-Brito, J. (2016). Use of the synthetic Oligo-pTa535 and Oligo-pAs1 probes for identification of *Hordeum chilense*-origin chromosomes in hexaploid *tritordeum*. *Genet. Resour. Crop. Evol.* 63, 945–951. doi: 10.1007/s10722-016-0402-3

Delibes, A., Del Moral, J., Martin-Sanchez, J. A., Mejias, A., Gallego, M., Casado, D., et al. (1997). Hessian fly-resistance gene transferred from Chromosome 4Mv of *Aegilops ventricosa* to *Triticum aestivum*. *Theor. Appl. Genet.* 94, 858–864. doi: 10.1007/s001220050487

Delibes, A., Lopez-Brafia, I., Mena, M., and Garcia-Olmedo, F. (1987). Genetic transfer of resistance to powdery mildew and of an associated biochemical marker from *Aegilops ventricosa* to hexaploid wheat. *Theor. Appl. Genet.* 73, 605–608. doi: 10.1007/BF00289201

Delibes, A., Romero, D., Aguaded, S., Duce, A., Mena, M., Lopez-Brana, I., et al. (1993). Resistance to cereal cyst nematode (*Heterodera avenae* Woll.) transferred from the wild grass *Aegilops ventricosa* to hexaploid wheat by a 'stepping stone' procedure. *Theor. Appl. Genet.* 87, 402–408. doi: 10.1007/BF01184930

Dosba, F., Doussinault, G., and Rivoal, R. (1978). "Extraction, identification and utilization of the addition lines T. aestivum–Ae.ventricosa," in *Proceedings of the 5th international wheat genetics symposium. Indian Society of Genetics and Plant Breeding*, ed. S. Ramanujam New Delhi, 332–337.

Dou, Q. W., Tanaka, H., Nakata, N., and Tsujimoto, H. (2006). Molecular cytogenetic analyses of hexaploid lines spontaneously appearing in octoploid *Triticale*. *Theor. Appl. Genet.* 114, 41–47. doi: 10.1007/s00122-006-0408-x

Doussinault, G., Delibes, A., Sanchez-Monge, R., and Garcia-Olmedo, F. (1983). Transfer of a dominant gene for resistance to eyespot disease from a wild grass to hexaploid wheat. *Nature* 303, 698–700. doi: 10.1038/303698a0

Dvorak, J., and Zhang, H. B. (1990). Variation in repeated nucleotide sequences sheds lights on the phylogeny of the wheat B and G genome. *Proc. Natl. Acad. Sci. U.S.A.* 87, 9640–9644. doi: 10.1073/pnas.87.24.9640

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Fores, R., Burlot, L., et al. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol.* 213, 1477–1486. doi: 10.1111/nph.14113

Friesen, N., Brandes, A., and Heslop-Harrison, J. S. (2001). Diversity, Origin, and distribution of. (retrotransposons)(gypsy and copia) in Conifers. *Mol. Biol. Evol.* 18, 1176–1188. doi: 10.1093/oxfordjournals.molbev.a003905

Fu, S. L., Chen, L., Wang, Y. Y., Li, M., Yang, Z. J., Qiu, L., et al. (2015). Oligonucleotide probes for ND-FISH analysis to identify Rye and wheat chromosomes. *Sci. Rep.* 5:10552. doi: 10.1038/srep10552

Guo, X. W., Bian, Y., Zhang, A., Zhang, H. K., Wang, B., Lv, R. L., et al. (2018). Transgenerationally precipitated meiotic chromosome instability fuels rapid karyotype evolution and phenotypic diversity in an artificially constructed allotetraploid wheat. *Mol. Biol. Evol.* 35, 1078–1091. doi: 10.1093/molbev/msy009

Han, F. P., Lamb, J. C., and Birchler, A. (2006). High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Pro. Natl. Acad. Sci. U.S.A.* 103, 3238–3243. doi: 10.1073/pnas.0509650103

Han, T. S., Wu, Q., Hou, X. H., Li, Z. W., Zhou, Y. P., Ge, S., et al. (2015). Frequent introgressions from diploid species contribute to the adaptation of the tetraploid Shepherd's Purse (*Capsella bursa-pastoris*). *Mol. Plant* 8, 427–438. doi: 10.1016/j.molp.2014.11.016

Hao, M., Luo, J. T., Zhang, L. Q., Yuan, Z. W., Yang, Y. W., Wu, M., et al. (2013). Production of hexaploid *triticale* by a synthetic hexaploid wheat-rye hybrid method. *Euphytica* 2013, 347–357. doi: 10.1007/s10681-013-0930-2

Huang, S. X., Sirikhachronkit, A., Su, X. J., Faris, J., Gill, B., Haselkorn, R., et al. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8133–8138. doi: 10.1073/pnas.072223799

Jahier, J., Abelard, P., Tanguy, A. M., Dedryver, F., Rivoal, R., Khathar, S., et al. (2001). The *Aegilops ventricosa* segment on chromosome 2AS of the wheat cultivar VPM1 carries the cereal cystnematode resistance gene *Cre5*. *Plant Breed.* 120, 125–128. doi: 10.1046/j.1439-0523.2001.00585.x

Kato, A., Lamb, J. C., and Birchler, J. A. (2004). Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13554–13559.

Kilian, B., Ozkan, H., Deusch, O., Effen, S., Brandolini, A., Kohl, J., et al. (2007). Independent wheat B and G genome origins in outcrossing *Aegilops* progenitors. *Mol. Biol. Evol.* 24, 217–227. doi: 10.1093/molbev/msl151

Lang, T., Li, G. R., Wang, H. J., Yu, Z. H., Chen, Q. H., Yang, E. N., et al. (2018). Physical location of tandem repeats in the wheat genome and application for chromosome identification. *Planta* 249, 663–675. doi: 10.1007/s00425-018-3033-4

Li, H., Guo, X. X., Wang, C. Y., and Ji, W. Q. (2015). Spontaneous and divergent hexaploid *Triticales* derived from common wheat ×rye by complete elimination of D-genome chromosomes. *PLoS One* 10:e0120421. doi: 10.1371/journal.pone.0120421

Li, H., Wang, Y. J., Guo, X. X., Du, Y. P., Wang, C. Y., and Ji, W. Q. (2016). Chromosomal structural changes and microsatellite variations in newly synthesized hexaploid wheat mediated by unreduced gemetes. *J. Genet.* 95, 819–830. doi: 10.1007/s12041-016-0704-4

Li, G., Gao, D., Zhang, H., Li, J., Wang, H., La, S., et al. (2016). Molecular cytogenetic characterization of *Dasypyrum breviaristatum* chromosomes in wheat background revealing the genomic divergence between *Dasypyrum* species. *Mol. Cytogenet.* 9:6. doi: 10.1186/s13039-016-0217-0

Liu, L. Q., Luo, Q. L., Teng, W., Li, B., Li, H. W., Li, Y. W., et al. (2018). Development of *Thinopyrum ponticum*-specifc molecular markers and FISH probes based on SLAF-seq technology. *Planta* 247, 1099–1108. doi: 10.1007/s00425-018-2845-6

Maìa, N. (1967). Obtention de blés tendres résistants au piétinverse par croisements interspécifiques blés × *Aegilops*. *C.R. Acad. Agric. Fr.* 53, 149–154.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 345:1250092. doi: 10.1126/science.1250092

McNeil, D., Lagudah, E. S., Hohmann, U., and Appels, R. (1994). Amplification of DNA sequences in wheat and its relatives: the Dgas44 and R350 families of repetitive sequences. *Genome* 37, 320–327. doi: 10.1139/g94-044

Mestiri, I., Chague, V., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., et al. (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. *New Phytol.* 186, 86–101. doi: 10.1111/j.1469-8137.2010.03186.x

Mirzaghaderi, G., and Mason, A. S. (2017). Revisiting pivotal-differential genome evolution in wheat. *Trends Plant Sci.* 22, 647–684. doi: 10.1016/j.tplants.2017.06.003

Ogbonnaya, F. C., Seah, S., Delibes, A., Jahier, J., Lopez-Brana, I., Eastwood, R. F., et al. (2001). Molecular-genetic characterization of a new nematode resistance gene in wheat. *Theor. Appl. Genet.* 102, 623–629. doi: 10.1007/s001220051689

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., et al. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant J.* 76, 1030–1044. doi: 10.1111/tpj.12366

Ren, T. H., He, M. J., Sun, Z. X., Tan, F. Q., Luo, P. G., Tang, Z. X., et al. (2019). The polymorphisms of Oligonucleotide probes in wheat cultivars determined by ND-FISH. *Molecules* 24:1126. doi: 10.3390/molecules24061126

Riley, R., and Kimber, G. (1961). Aneuploids and the cytogenetic structure of wheat varietal populations. *Heredity* 16, 275–290.

Senerchia, N., Felber, F., and Parisod, C. (2014). Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. *New Phytol.* 202, 975–985. doi: 10.1111/nph.12731

Su, Y. R., Zhang, D. L., Li, Y., and Li, S. P. (2015). Nonhomologous Chromosome pairing in *Aegilops*-Secale Hybrids. *Cytogenet. Genome Res.* 147, 268–273. doi: 10.1159/000444435

Tang, S., Tang, Z., Qiu, L., Yang, Z., Li, G., Lang, T., et al. (2018). Developing new Oligo probes to distinguish specific chromosomal segments and the A, B, D genomes of wheat (*Triticum aestivum* L.) using ND-FISH. *Front. Plant Sci.* 9:1104. doi: 10.3389/fpls.2018.01104

Tang, S. Y., Qiu, L., Xiao, Z. Q., Fu, S. L., and Tang, Z. X. (2016). New Oligonucleotide probes for ND-FISH analysis to identify barley chromosomes and to investigate polymorphisms of wheat chromosomes. *Genes* 7:118. doi: 10.3390/genes7120118

Tang, Z. X., Yang, Z. J., and Fu, S. L. (2014). Oligonucleotides replacing the roles of repetitive sequences pAs1, pSc119.2, *pTa*-535, pTa71, CCS1, and pAWRC.1 for FISH analysis. *J. Appl. Genet.* 55, 313–318. doi: 10.1007/s1335

Tanguy, A.-M., Coriton, O., Abelard, P., Dedryver, F., and Jahier, J. (2005). Structure of *Aegilops ventricosa* chromosome 6Nv, the donor of wheat genes *Yr17*. *Lr37*, *Sr38*, and *Cre5*. *Genome* 48, 541–546. doi: 10.1139/g05-001

Tiwari, V. K., Rawat, N., Neelam, K., Kumar, S., Randhawa, G. S., and Dhaliwal, H. S. (2010). Random chromosome elimination in synthetic *Triticum-Aegilops* amphiploids leads to development of a stable partial amphiploid with high grain micro- and macronutrient content and powdery mildew resistance. *Genome* 53, 1053–1065. doi: 10.1139/G10-083

Xie, Q., Kang, H. Y., Tao, S., Sparkes, D. L., Fan, X., Cui, Z., et al. (2012). Wheat lines derived from trigeneric hybrids of wheat-rye-*Psathyrostachys huashanica*, the potentical resources for grain weight improvement. *Aust. J. Crop Sci.* 6, 1550–1557.

Zhang, D. L., He, J., Huang, L. Y., Zhang, C. C., Zhou, Y., Su, Y. R., et al. (2017). An advanced backcross population through synthetic octaploid wheat as a "bridge":

development and QTL detection for seed dormancy. *Front.Plant Sci.* 8:2123. doi: 10.3389/fpls.2017.02123

Zhang, H. K., Bian, Y., Gou, X. W., Zhu, B., Xu, C. M., Qi, B., et al. (2013a). Persistenet whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3447–3452. doi: 10.1073/pnas.1300153110

Zhang, H. K., Bian, Y., Gou, X. W., Dong, Y. Z., Rustgi, S., Zhang, B. J., et al. (2013b). Intrinsic karyotype stability and gene copy number variations may have laid the foundation for tetraploid wheat formation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19466–19471. doi: 10.1073/pnas.1319598110

Zhang, J., Jiang, Y., Guo, Y. L., Li, G. R., Yang, Z. J., Xu, D. L., et al. (2015). Identification of novel chromosomal aberrations induced by 60Co-γ-Irradiation in wheat-*Dasypyrum villosum* Lines. *Int J Mol Sci.* 16, 29787–29796. doi: 10.3390/ijms161226134

Zhang, J., Jiang, Y., Wang, Y., Guo, Y., Long, H., Deng, G., et al. (2018). Molecular markers and cytogenetics to characterize a wheat-*Dasypyrum villosum* 3V (3D) substitution line conferring resistance to stripe rust. *PLoS One* 13:e0202033. doi: 10.1371/journal.pone.0202033

Zhang, D. L., Zhou, Y., Zhao, X. P., Lv, L. L., Zhang, C. C., Li, J. H., et al. (2018). Development and utilization of introgression lines using synthetic octaploid wheat (*Aegilops tauschii* × hexaploid Wheat) as donor. *Front. Plant Sci.* 9:1113. doi: 10.3389/fpls.2018.01113

Zhao, L. B., Ning, S. Z., Yi, Y. J., Zhang, L. Q., Yuan, Z. W., Zeng, Y. L., et al. (2018). Fluorescence in situ hybridization karyotyping reveals the presence of two distinct genomes in the taxon *Aegilops tauschi*. *BMC Genomics* 19:3. doi: 10.1186/s12864-017-4384-0

Zhao, N., Xu, L. Y., Zhu, B., Li, M. J., Zhang, H. K., Qi, B., et al. (2011). Chromosomal and genome-wide molecular changes associated with initial stages of allohexaploidization in wheat can be transit and incidental. *Genome* 54, 692–699. doi: 10.1139/G11-028

Zhou, J. P., Zhang, H. Y., Yang, Z. J., Li, G. R., Hu, L. J., Lei, M. P., et al. (2012). Characterization of a new T2DS.2DL-?R translocation triticale ZH-1 with multiple resistances to diseases. *Genet. Resour. Crop. Evol.* 59, 1161–1168. doi: 10.1007/s10722-011-9751-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/fgene.2020.00455

LuluDB—The Database Created Based on Small RNA, Transcriptome, and Degradome Sequencing Shows the Wide Landscape of Non-coding and Coding RNA in Yellow Lupine (*Lupinus luteus* L.) Flowers and Pods

Paulina Glazinska[1,2], Milena Kulasek[1,2], Wojciech Glinkowski[1,2], Marta Wysocka[3] and Jan Grzegorz Kosiński[3]*

[1] *Department of Plant Physiology and Biotechnology, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University, Torun, Poland,* [2] *Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, Torun, Poland,* [3] *Department of Computational Biology, Faculty of Biology, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland*

OPEN ACCESS

Edited by:
Wenqin Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Changjiang Yu,
Qingdao Institute of Bioenergy and Bioprocess Technology (CAS), China
Yong2017 Zhou,
King Abdullah University of Science and Technology, Saudi Arabia

****Correspondence:***
Paulina Glazinska
paulina.glazinska@umk.pl

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *10 February 2020*
Accepted: *14 April 2020*
Published: *15 May 2020*

Citation:
Glazinska P, Kulasek M, Glinkowski W, Wysocka M and Kosiński JG (2020) LuluDB—The Database Created Based on Small RNA, Transcriptome, and Degradome Sequencing Shows the Wide Landscape of Non-coding and Coding RNA in Yellow Lupine (Lupinus luteus L.) Flowers and Pods. Front. Genet. 11:455. doi: 10.3389/fgene.2020.00455

Yellow lupine (*Lupinus luteus* L.) belongs to a legume family that benefits from symbiosis with nitrogen-fixing bacteria. Its seeds are rich in protein, which makes it a valuable food source for animals and humans. Yellow lupine is also the model plant for basic research on nodulation or abscission of organs. Nevertheless, the knowledge about the molecular regulatory mechanisms of its generative development is still incomplete. The RNA-Seq technique is becoming more prominent in high-throughput identification and expression profiling of both coding and non-coding RNA sequences. However, the huge amount of data generated with this method may discourage other scientific groups from making full use of them. To overcome this inconvenience, we have created a database containing analysis-ready information about non-coding and coding *L. luteus* RNA sequences (LuluDB). LuluDB was created on the basis of RNA-Seq analysis of small RNA, transcriptome, and degradome libraries obtained from yellow lupine cv. Taper flowers, pod walls, and seeds in various stages of development, flower pedicels, and pods undergoing abscission or maintained on the plant. It contains sequences of miRNAs and phased siRNAs identified in *L. luteus*, information about their expression in individual samples, and their target sequences. LuluDB also contains identified lncRNAs and protein-coding RNA sequences with their organ expression and annotations to widely used databases like GO, KEGG, NCBI, Rfam, Pfam, etc. The database also provides sequence homology search by BLAST using, e.g., an unknown sequence as a query. To present the full capabilities offered by our database, we performed a case study concerning transcripts annotated as *DCL 1–4* (*DICER LIKE 1–4*) homologs involved in small non-coding RNA biogenesis and identified miRNAs that most likely regulate *DCL1* and *DCL2* expression in yellow lupine. LuluDB is available at http://luluseqdb.umk.pl/basic/web/index.php.

Keywords: yellow lupine, RNA-Seq, database, flower, pod, miRNA, siRNA, long ncRNA

INTRODUCTION

Yellow lupine (*Lupinus luteus* L.) belongs to a legume family that benefits from symbiosis with nitrogen-fixing bacteria. Seeds of this species are rich in proteins that constitute up to 40% of their dry mass (Ogura et al., 2014). Additionally, years of research and selective breeding have led to the development of alkaloid-free "sweet" cultivars. All these traits make lupine seeds a valuable food source for animals and humans primarily in climatic conditions unfavorable for soybean cultivation (Musco et al., 2017).

The main constraint on a large-scale cultivation of yellow lupine comes from its excessive shedding of generative organs, which contributes to significant yield losses. Therefore, current research focuses on the development of varieties of yellow lupine and cultivation conditions that would prevent massive flower and pod dropping, consequently stabilizing the yield in various environmental conditions (Lucas et al., 2015). Besides its practical significance, yellow lupine is also an excellent model plant for basic research on nodulation (Frankowski et al., 2015) or abscission of generative organs (Glazinska et al., 2017, 2019).

Advances in high-throughput techniques have found new opportunities for deeper exploration of complex nets of factors that regulate biological processes. However, it generates tremendous amount of data, which is impossible to analyze without powerful computers and programming skills. For example, in databases like SRA NCBI, only raw data are deposited, which makes the information unavailable to a wider scientific audience. Due to the current trend in analyzing big amounts of biological data in evolutionary context, it is of great importance to provide the users with the most convenient way possible. One of the best solutions includes the creation of a database with user-friendly interface and downloadable data in the form of analysis-ready tables.

Exemplary databases of this type for other plant species usually contain data on one type of RNA, either encoding proteins (Kawahara et al., 2016; Robinson et al., 2018) or non-coding (Liu et al., 2012; Gupta et al., 2018). Based on their length, non-coding RNAs (ncRNA) are classified into short (<200 nt) and long (over 200 nt) categories (Liu et al., 2015). Short ncRNAs (sncRNA) are represented by miRNA (micro RNAs) and phased siRNA (phased, secondary, small interfering RNAs originally designated as trans-acting small interfering RNAs or ta-siRNAs) (Axtell, 2013a; Fei et al., 2013). They are involved in post-transcriptional control of their target gene activity in the process of RNA interference. Short ncRNAs binding to specific mRNA on the principle of complementarity leads to either its cleavage within the bound sequence or inhibition of its translation (Bartel, 2004; Vaucheret, 2006). Long ncRNAs (lncRNAs) were shown to be potent *cis*- and *trans*-regulators of gene transcription, and can act as (i) scaffolds for chromatin-modifying complexes, (ii) decoy for splicing factors, or (iii) competitors for miRNA binding sites (Marchese et al., 2017). Regarding legumes, a LegumeIP 2.0 database is available, containing data on genomic and protein-coding sequences for six legume species: *Medicago truncatula* (lucerne), *Glycine max* (soybean), *Lotus japonicus* (birdsfoot trefoil), *Phaseolus vulgaris* (common bean), *Cicer arietinum* (chickpea), and *Cajanus cajan* (pigeon pea) and two outgroup reference species: *Arabidopsis thaliana* and *Populus trichocarpa* (Li et al., 2016). Other examples are SoyNET (Kim et al., 2017) and SoyKB (Joshi et al., 2017) for soybean. The latter is the most extensive one, which also contains information on sRNA sequences, however, only for *Glycine*.

In case of *Lupinus luteus* database (LuluDB), our aim was an integration of our RNA-Seq data for yellow lupine protein-coding RNA and ncRNA sequences in one place. Besides protein-coding transcripts, LuluDB contains information about known and novel miRNAs, siRNAs and their target transcripts, as well as lncRNAs. LuluDB provides information about sequences, their accumulation in generative organs and identified target transcripts for sRNAs. It is possible to view the database by scrolling the interactive list of elements, search it by sequence ID, or query it by homological sequence using built-in BLASTn. LuluDB will popularize the genetic research on this important crop plant and support the research on universal regulatory mechanisms of plant development mediated by ncRNAs.

Detailed analysis of the data concerning miRNA and siRNA in yellow lupine flowers has been already published (Glazinska et al., 2019), while similar analysis conducted for pods is presented here for the first time. Similarly, data on putative lncRNA in lupine have not been published yet. A short presentation of these results will be presented here together with a description of the capabilities of the database. In addition, to prove the usefulness of the database, we present a case study of transcripts identified as *DCL1–4* (*DICER LIKE 1–4*) homologs involved in the small ncRNA biogenesis process (Fukudome and Fukuhara, 2017) and identified miRNAs that most likely regulate *DCL1* and *DCL2* expression in yellow lupine. The database is available at http://luluseqdb.umk.pl/basic/web/index.php.

DESCRIPTION OF UTILITY AND DISCUSSION

Data Sources and Generation in LuluDB

LuluDB was created on the basis of NGS sequencing analysis of sRNA, transcriptomes, and degradome libraries obtained from generative organs of yellow lupine cv. Taper: flowers in various stages of development, developing pod walls and seeds, flower pedicels, and pods undergoing abscission and control ones. Through this experimental design, we aimed at examining global changes in expression during flower development, and wanted to determine the differences in their development depending on the location in the inflorescence, which is associated with the tendency to fall off/transform into pods (van Steveninck, 1959). Therefore, flowers from the highest whorls of inflorescences (of which 100% undergo abscission) (UF) and from the lowest (100% binds pods) (LF) were collected separately in four developmental variants. The first stage consisted of closed, yellowing, elongating buds with closed anthers, and the second stage consisted of closed buds with yellow petals and open anthers; during the third stage, there were flowers in full anthesis, with visible pollen on the stigma, and the fourth stage comprised open flowers with aging anthers, and no trace of sticky pollen, but with yellow petals retaining turgor. The ovules of the lower fourth stage flowers were enlarged. To explore the landscape of RNA expression during the abscission process, we also collected pedicels of flowers at stage 3 from the lower whorls (FPNAB) and of flowers with symptoms of abscission and senescence (FPAB). Due to the fact that lupine pods fall off at the initial stages of development, we also sampled young green pods without active abscission zone (PNAB) and pods of the same age with visible symptoms of abscission (PAB). Yellow lupine is a plant valued for its seeds (Lucas et al., 2015). Understanding the global changes in gene expression during the development of pods of this plant may become a valuable contribution to various studies aiming at improving crop yields. The pod RNA libraries were constructed separately from seeds (S) and pod walls (W) collected in eight time points, and then combined into three variants, where the first one (P1) included early developmental stages characterized by intensive growth, the second one (P2) wherein seed filling occurred, and the third one (P3) where the filling ended and the pods started to ripen and desiccate. The detailed list of samples is presented in the **Table 1**.

After the sequencing and preliminary data analysis, the data concerning sequences of identified coding RNAs and ncRNAs were first deposited in the raw form in NCBI SRA database and then analysis-ready data were uploaded into the LuluDB database. **Table 1** and **Supplementary Table 1** present details of the data deposited in NCBI SRA and the LuluDB.

ncRNAs in LuluDB

The database contains sequences of 456 known and 32 novel miRNAs, as well as 318 phased siRNAs identified in yellow lupine along with information about their expression and target transcripts. In our previous paper (Glazinska et al., 2019) in which part of the data described here supported the evidence that sRNAs are involved in yellow lupine flower development and abscission, each miRNA received a unique ID number on a slightly different principle than it is presented in LuluDB. Namely, known miRNAs [i.e., having identical hits in miRBase (Kozomara et al., 2019)] were assigned IDs from Ll-miR1 to Ll-miR456, and the numbering of novel miRNAs [identified with ShortStack (Axtell, 2013b)] started from the beginning, with the "n" inset before the number (e.g., Ll-miRn22). As the small RNA-Seq technology becomes more available, we predict an increase in deposition of new miRNA sequences to miRBase, and in the future, it might be discovered that some of the lupine miRNAs currently considered as novel have homologs in other plant species, and such numeration would be misleading. This is why in LuluDB the numbering within novel miRNA IDs continues after the last known one, from 457 up. Besides the identification of novel miRNAs, ShortStack was used to identify small RNA cut in phase from longer precursors (phased siRNA).

Expression of small RNAs in individual samples is stated in RPM (reads per million). For both miRNAs and siRNAs, potential target transcripts were identified by degradome data analysis carried out with CleaveLand4 (Addo-Quaye et al., 2009) and by additional *in silico* analysis with psRNATarget toolkit (Dai et al., 2018) in order to predict targets that are not only cleaved, but may also be suppressed in other modes of sRNA action.

We have identified lncRNAs by performing BLASTn search within CantataDB (Szcześniak et al., 2019) in which *G. max* lncRNAs were queried by transcripts obtained in our experiment. As a result, 31,718 lncRNA sequences homologous to those in soybean were found and deposited in LuluDB.

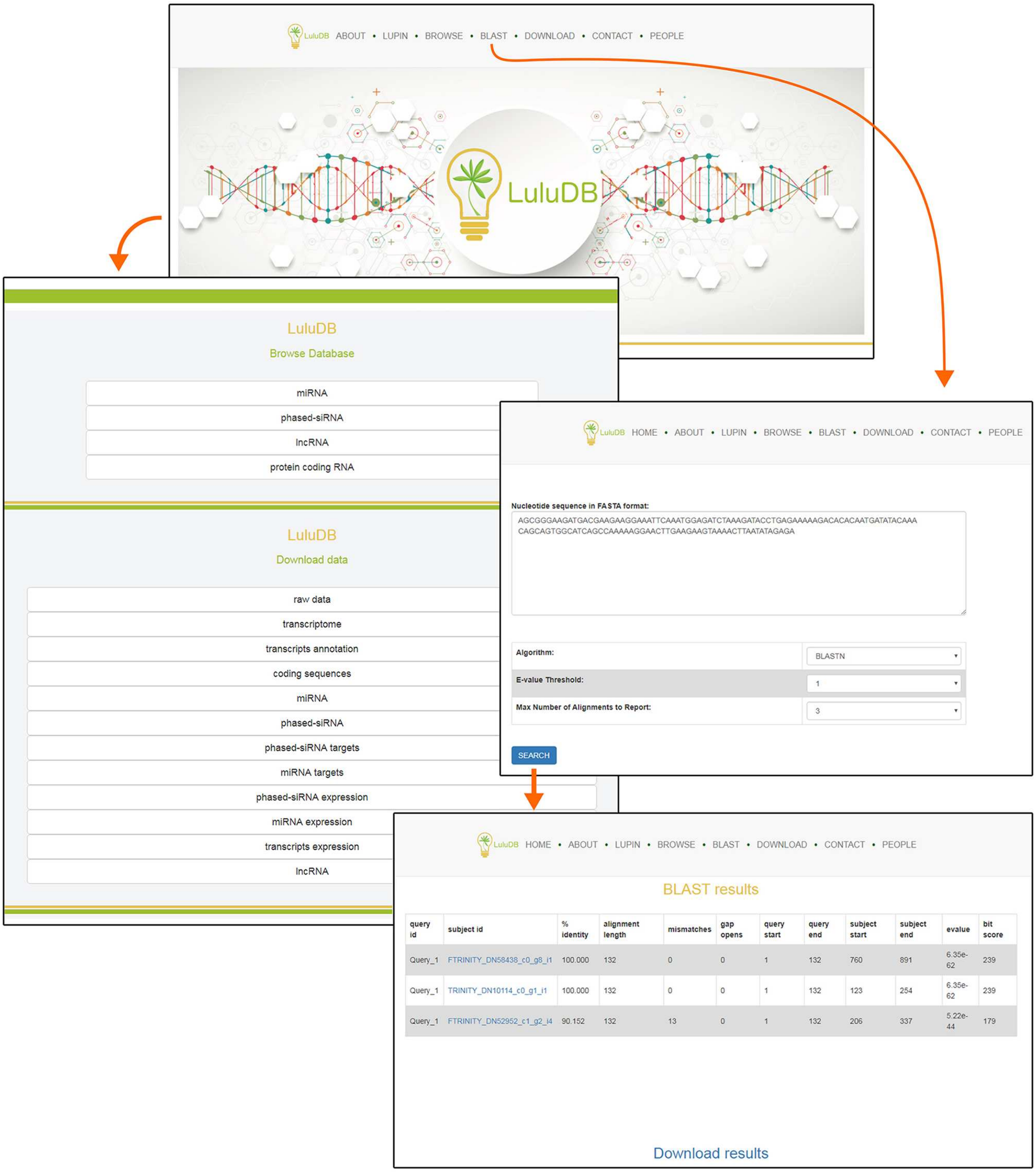
Coding RNAs in LuluDB

LuluDB contains 267,349 protein-coding RNA sequences with annotations to commonly used databases: Blastp, Blastx, Eggnog,

TABLE 1 | List of samples deposited to date in the LuluDB database.

Alias	Sample name	Description	RNA-Seq	Small RNA-Seq	Degradome	References
UF1	Upper flowers stage 1	Flowers from upper part of raceme in stage 1	•	•		(Glazinska et al., 2019), this study
UF2	Upper flowers stage 2	Flowers from upper part of raceme in stage 2	•	•		(Glazinska et al., 2019), this study
UF3	Upper flowers stage 3	Flowers from upper part of raceme in stage 3	•	•	•	(Glazinska et al., 2019), this study
UF4	Upper flowers stage 4	Flowers from upper part of raceme in stage 4	•	•		(Glazinska et al., 2019), this study
LF1	Lower flowers stage 1	Flowers from lower part of raceme in stage 1	•	•		(Glazinska et al., 2019), this study
LF2	Lower flowers stage 2	Flowers from lower part of raceme in stage 2	•	•		(Glazinska et al., 2019), this study
LF3	Lower flowers stage 3	Flowers from lower part of raceme in stage 3	•	•	•	(Glazinska et al., 2019), this study
LF4	Lower flowers stage 4	Flowers from lower part of raceme in stage 4	•	•		(Glazinska et al., 2019), this study
FPNAB	Flower pedicels non-abscissing	Pedicels of non-abscissing flowers	•	•		(Glazinska et al., 2019), this study
FPAB	Flower pedicels abscissing	Pedicels of abscissing flowers	•	•		(Glazinska et al., 2019), this study
PW1	Pod walls stage 1	Pod walls in early stage of development	•	•		This study
PW2	Pod walls stage 2	Pod walls in middle stage of development	•	•		This study
PW3	Pod walls stage 3	Pod walls in late stage of development	•	•	•	This study
PS1	Pod seeds stage 1	Seeds in early stage of development	•	•		This study
PS2	Pod seeds stage 2	Seeds in middle stage of development	•	•		This study
PS3	Pod seeds stage 3	Seeds in late stage of development	•	•	•	This study
PNAB	Pods non-abscissing	Non-abscissing pods	•	•		This study
PAB	Pods abscised	Abscissing pods	•	•		This study

KEGG, CantataDB, miRBase, NCBI protein, Pfam, Rfam, and GO. Because the reference yellow lupine genome sequencing is still in progress (Iqbal et al., 2020), the transcripts were assembled *de novo*. This assembly was carried out separately for libraries derived from flowers (already published in Glazinska et al., 2019) and pods with Trinity toolkit, which assigned an ID for each transcript (e.g., TRINITY_DN10038_c0_g1_i1) within each batch separately. This created the risk that completely


LuluDB
ABOUT • LUPIN • BROWSE • BLAST • DOWNLOAD • CONTACT • PEOPLE
LuluDB
LuluDB HOME • ABOUT • LUPIN • BROWSE • BLAST • DOWNLOAD • CONTACT • PEOPLE
SEARCH
LuluDB HOME • ABOUT • LUPIN • BROWSE • BLAST • DOWNLOAD • CONTACT • PEOPLE

LuluDB

Browse Database

- miRNA
- phased-siRNA
- lncRNA
- protein coding RNA

LuluDB

Download data

- raw data
- transcriptome
- transcripts annotation
- coding sequences
- miRNA
- phased-siRNA
- phased-siRNA targets
- miRNA targets
- phased-siRNA expression
- miRNA expression
- transcripts expression
- lncRNA

Nucleotide sequence in FASTA format:

```
AGCGGGAAGATGACGAAGAAGGAAATTCAAATGGAGATCTAAAGATACCTGAGAAAAAGACACACAATGATATACAAA
CAGCAGTGGCATCAGCCAAAAAGGAACTTGAAGAAGTAAAACTTAATATAGAGA
```

Algorithm:	BLASTN
E-value Threshold:	1
Max Number of Alignments to Report:	3

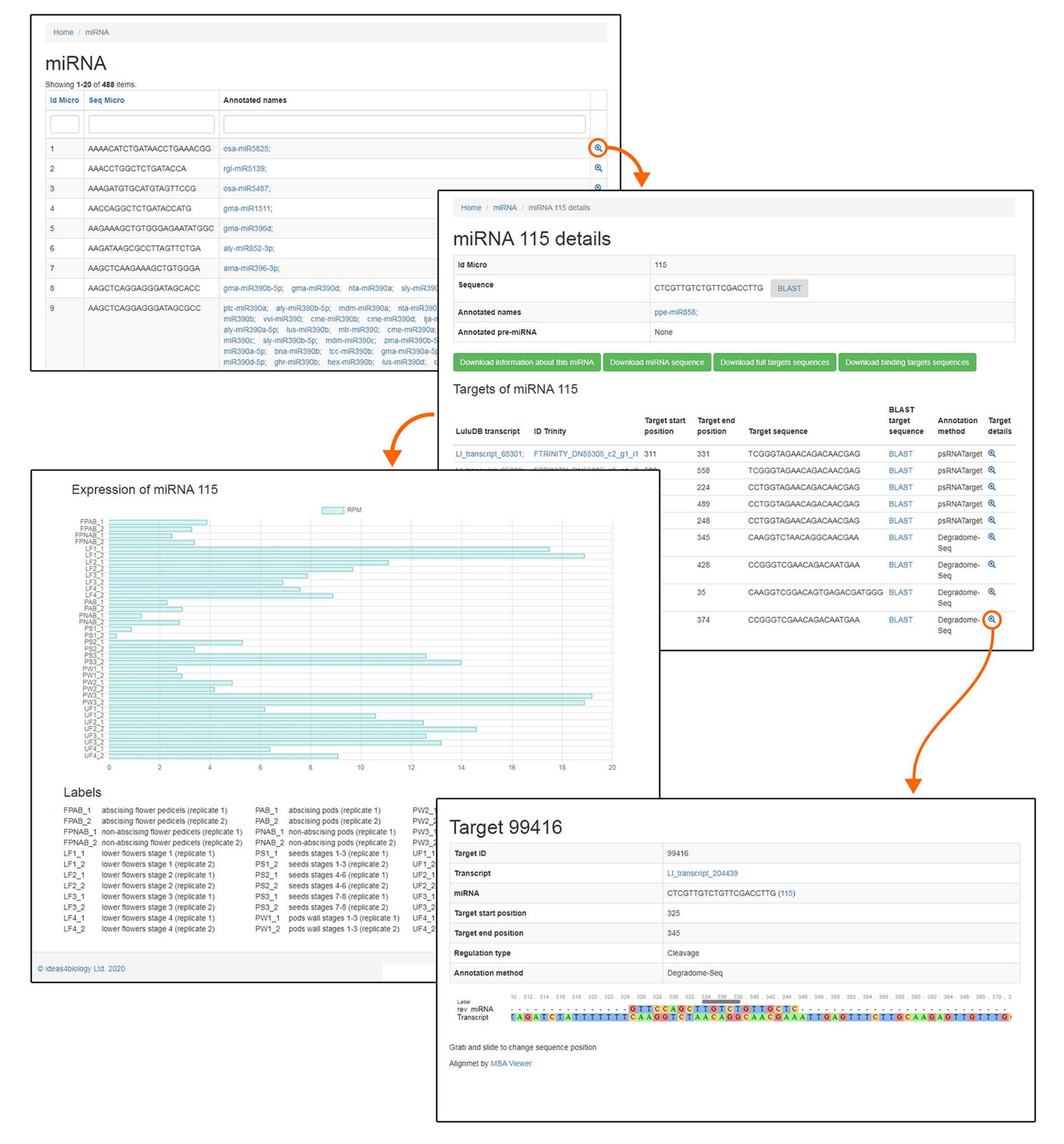
BLAST results

query id	subject id	% identity	alignment length	mismatches	gap opens	query start	query end	subject start	subject end	evalue	bit score
Query_1	FTRINITY_DN58438_c0_g8_i1	100.000	132	0	0	1	132	760	891	6.35e-62	239
Query_1	TRINITY_DN10114_c0_g1_i1	100.000	132	0	0	1	132	123	254	6.35e-62	239
Query_1	FTRINITY_DN52952_c1_g2_i4	90.152	132	13	0	1	132	206	337	5.22e-44	179

Download results

FIGURE 1 | Screenshot of LuluDB home page and of the interface to submit BLAST searches.

dissimilar transcripts in flowers and pods could have the same ID. We fixed this in three ways: (i) by providing each transcript with information about its origin (flowers or pods), (ii) by adding the "F" prefix to TRINITY ID in flower dataset (e.g., FTRINITY_DN53848_c2_g1_i5), and (iii) by assigning additional ID for the database (e.g., LI_transcript_534367). All of the assembled transcripts were clustered and, within each cluster, they were assigned "Gene" name: (e.g., LI_gene_11901). Analysis indicates that the majority, but not all of the transcripts, have their near-identical homologs in both types of organs



Home / miRNA

miRNA

Showing **1-20** of **488** items.

Id Micro	Seq Micro	Annotated names	
1	AAAACATCTGATAACCTGAAACGG	osa-miR5825;	🔍
2	AAACCTGGCTCTGATACCA	rgl-miR5139;	🔍
3	AAAGATGTGCATGTAGTTCCG	osa-miR5487;	🔍
4	AACCAGGCTCTGATACCATG	gma-miR1511;	
5	AAGAAAGCTGTGGGAGAATATGGC	gma-miR396d;	
6	AAGATAAGCGCCTTAGTTCTGA	aly-miR852-3p;	
7	AAGCTCAAGAAAGCTGTGGGA	ama-miR396-3p;	
8	AAGCTCAGGAGGGATAGCACC	gma-miR390b-5p; gma-miR390d; nta-miR390a; sly-miR390	
9	AAGCTCAGGAGGGATAGCGCC	ptc-miR390a; aly-miR390b-5p; mdm-miR390a; nta-miR390 miR390b; vvi-miR390; cme-miR390b; cme-miR390d; lja-m aly-miR390a-5p; lus-miR390b; mtr-miR390; cme-miR390a; miR390c; sly-miR390b-5p; mdm-miR390c; zma-miR390b-5 miR390a-5p; bna-miR390b; tcc-miR390b; gma-miR390a-5p miR390d-5p; ghr-miR390b; hex-miR390b; lus-miR390d; c	

Home / miRNA / miRNA 115 details

miRNA 115 details

Id Micro	115
Sequence	CTCGTTGTCTGTTCGACCTTG BLAST
Annotated names	ppe-miR858;
Annotated pre-miRNA	None

Download information about this miRNA | Download miRNA sequence | Download full targets sequences | Download binding targets sequences

Targets of miRNA 115

LuluDB transcript	ID Trinity	Target start position	Target end position	Target sequence	BLAST target sequence	Annotation method	Target details
LI_transcript_65301;	FTRINITY_DN55305_c2_g1_i1	311	331	TCGGGTAGAACAGACAACGAG	BLAST	psRNATarget	🔍
			558	TCGGGTAGAACAGACAACGAG	BLAST	psRNATarget	🔍
			224	CCTGGTAGAACAGACAACGAG	BLAST	psRNATarget	🔍
			489	CCTGGTAGAACAGACAACGAG	BLAST	psRNATarget	🔍
			248	CCTGGTAGAACAGACAACGAG	BLAST	psRNATarget	🔍
			345	CAAGGTCTAACAGGCAACGAA	BLAST	Degradome-Seq	🔍
			428	CCGGGTCGAACAGACAATGAA	BLAST	Degradome-Seq	🔍
			35	CAAGGTCGGACAGTGAGACGATGGG	BLAST	Degradome-Seq	🔍
			374	CCGGGTCGAACAGACAATGAA	BLAST	Degradome-Seq	🔍

Expression of miRNA 115

Labels

FPAB_1	abscising flower pedicels (replicate 1)	PAB_1	abscising pods (replicate 1)	PW2_1
FPAB_2	abscising flower pedicels (replicate 2)	PAB_2	abscising pods (replicate 2)	PW2_2
FPNAB_1	non-abscising flower pedicels (replicate 1)	PNAB_1	non-abscising pods (replicate 1)	PW3_1
FPNAB_2	non-abscising flower pedicels (replicate 2)	PNAB_2	non-abscising pods (replicate 2)	PW3_2
LF1_1	lower flowers stage 1 (replicate 1)	PS1_1	seeds stages 1-3 (replicate 1)	UF1_1
LF1_2	lower flowers stage 1 (replicate 2)	PS1_2	seeds stages 1-3 (replicate 2)	UF1_2
LF2_1	lower flowers stage 2 (replicate 1)	PS2_1	seeds stages 4-6 (replicate 1)	UF2_1
LF2_2	lower flowers stage 2 (replicate 2)	PS2_2	seeds stages 4-6 (replicate 2)	UF2_2
LF3_1	lower flowers stage 3 (replicate 1)	PS3_1	seeds stages 7-8 (replicate 1)	UF3_1
LF3_2	lower flowers stage 3 (replicate 2)	PS3_2	seeds stages 7-8 (replicate 2)	UF3_2
LF4_1	lower flowers stage 4 (replicate 1)	PW1_1	pods wall stages 1-3 (replicate 1)	UF4_1
LF4_2	lower flowers stage 4 (replicate 2)	PW1_2	pods wall stages 1-3 (replicate 2)	UF4_2

© ideas4biology Ltd. 2020

Target 99416

Target ID	99416
Transcript	LI_transcript_204439
miRNA	CTCGTTGTCTGTTCGACCTTG (115)
Target start position	325
Target end position	345
Regulation type	Cleavage
Annotation method	Degradome-Seq

```
           10 . 312 . 314 . 316 . 318 . 320 . 322 . 324 . 326 . 328 . 330 . 332 . 334 . 336 . 338 . 340 . 342 . 344 . 346 . 348 . 350 . 352 . 354 . 356 . 358 . 360 . 362 . 364 . 366 . 368 . 370 . 3
Label
rev miRNA  . . . . . . . . . . . . . . . G T T C C A G C T T G T C T G T T G C T C . . . . . . . . . . . . . . . . . . . . . . . . . .
Transcript T A G A T C T A T T T T T T T T C A A G G T C T A A C A G G C A A C G A A A T T G A G T T T C T T G C A A G A G T T G T T T G
```

Grab and slide to change sequence position

Alignmet by MSA Viewer

FIGURE 2 | Screenshot of LuluDB page concerning example miRNA.

(**Supplementary Figure 1**). Minor differences may be caused either by assembly errors or the specificity of transcript processing (such as alternative splicing).

The expression levels in FPKM unit (fragments per kilobase of exon per million fragments mapped) are shown only for the relevant libraries.

Database Organization

You can easily navigate to major database components from the top of the home page (**Figure 1**). The selection includes general information (About), information on *L. luteus* (Lupin), browsers for various types of deposited sequences (Browse), integrated BLAST search tool for user sequences (BLAST), data download links (Download), contact information to corresponding author (Contact), and information on everyone involved in the creation of LuluDB (People) (**Figure 1**).

One of the most crucial elements of the home page is the Browse section. This page contains links to various parts of the database, such as miRNA, phased siRNA, lncRNA, as well as protein-coding RNAs (**Figure 1**). A list of identified miRNAs can be found in the miRNA section and searched by sequence ID, RNA sequence, or miRBase annotation (**Figure 2**), and the annotations also serve as links to relevant miRBase entries. The magnifying glass icon in the last column leads to the details site where you can find information about its Id Micro, Sequence, Annotated names, and Annotated pre-miRNA. If target transcripts were identified for this particular miRNA, a handy table containing detailed information about it is displayed lower in this page. Below, expression of the miRNA is plotted for each library as a bar plot. It is possible to download all data contained on this page by clicking clearly described buttons. In the “details” section, the BLASTn tool (redirect to BLASTn with already loaded sequence of either miRNA and target gene) allows for a quick analysis.

The phased siRNA section is structured in similar manner (**Supplementary Figure 2A**).

On the lncRNA main page, there is a list of transcripts identified as lncRNAs, composed of ID from Cantata, LuluDB transcript ID, and Trinity Id (**Supplementary Figure 2B**). Clicking the magnifying glass icon in the last column leads to a page with details on the given transcript: internal LuluDB ID, ID from Cantata with hyperlink to that database, LuluDB transcript ID with internal hyperlink to more details about this transcript, including its expression, and its sequence. As in previous cases, all this information can be downloaded. This part of the database is the least extensive due to the limited ways of analyzing lncRNAs.

In the protein-coding transcript section, the list of transcripts can be searched by TRINITY ID, LuluDB ID, or ORF type, which can be: “complete,” “internal,” “3prime_partial,” or “5prime_partial” or annotations to various databases (**Supplementary Figure 3**). Under the magnifying glass icon, there is a link to a page with details: LuluDB ID, CDS coordinates, predicted amino acid sequence, and details about annotation to following databases: Blastp, Blastx, Eggnog, Kegg, CantataDB, miRBase, NCBI protein, Pfam, Rfam, GO, and graphical representation of ORF. All of the data are available for download. Under LuluDB ID, there is a hyperlink to page with more information about this transcript, including its expression and annotation details. **Table 2** illustrates the number of transcripts identified in yellow lupine annotated to public databases.

TABLE 2 | Summary of protein-coding transcripts deposited to date in LuluDB annotated in various open access databases.

Public database	No. of annotated unigenes
RFAM	4,433
PFAM	198,225
CantataDB	31,718
Nr	288,854
SwissProt	216,711
KEGG	247,375
GO	534,413
miRBase	2,565

Validation of sRNA and RNA Sequencing Results Deposited in LuluDB

Results of sRNA and RNA deep sequencing were validated using qPCR technique. Validation for data concerning flowers was already presented in our recent paper (Glazinska et al., 2019) and the same approach was used to validate the rest of the results. Regarding sRNA-Seq, 7 miRNA, and 2 siRNA sequences were selected and subjected to stem-loop RT-qPCR technique (Kramer, 2011; Varkonyi-Gasic and Hellens, 2011), while in the case of RNA-Seq, the standard qPCR reaction was carried out for eight transcripts. **Figure 3** shows the plotted $\log_2$ ratio of fold changes (FC) calculated from qPCR against $\log_2$ FC of the sRNA-Seq (**Figure 3A**) or RNA-Seq data (**Figure 3B**). **Figure 3C** shows the validation data for nine abovementioned small RNAs and **Supplementary Figure 4** displays the validation data of eight transcripts used for validation for homologous transcripts found in both flowers and pods. The R^2 and Spearman’s rank correlation coefficient (rho) are satisfactory and confirm the linearity of the relationship between qPCR and sRNA-Seq or RNA-Seq data (**Figures 3A,B**), and similarity in expression patterns measured with NGS and qPCR (**Figure 3C** and **Supplementary Figure 4**) supports the validity of our RNA-Seq.

CASE STUDY

Identification of Homologs of DCL Family Members in Yellow Lupine

In order to show the functionality of LuluDB and present an exemplary analysis pipeline, we used the database interface to identify transcripts encoding homologs of *A. thaliana* DCL1, 2, 3, and 4 involved in small ncRNA biogenesis (Fukudome and Fukuhara, 2017) and miRNAs that post-transcriptionally regulate their expression in yellow lupine.

Firstly, we downloaded the CDS sequences of selected *A. thaliana DCL* genes from NCBI (NM_001197952, NM_001202869.2, NM_001161191.3, NM_001203419.2) (**Supplementary Table 2**). Then, we queried LuluDB with these sequences using local BLASTn tool with the

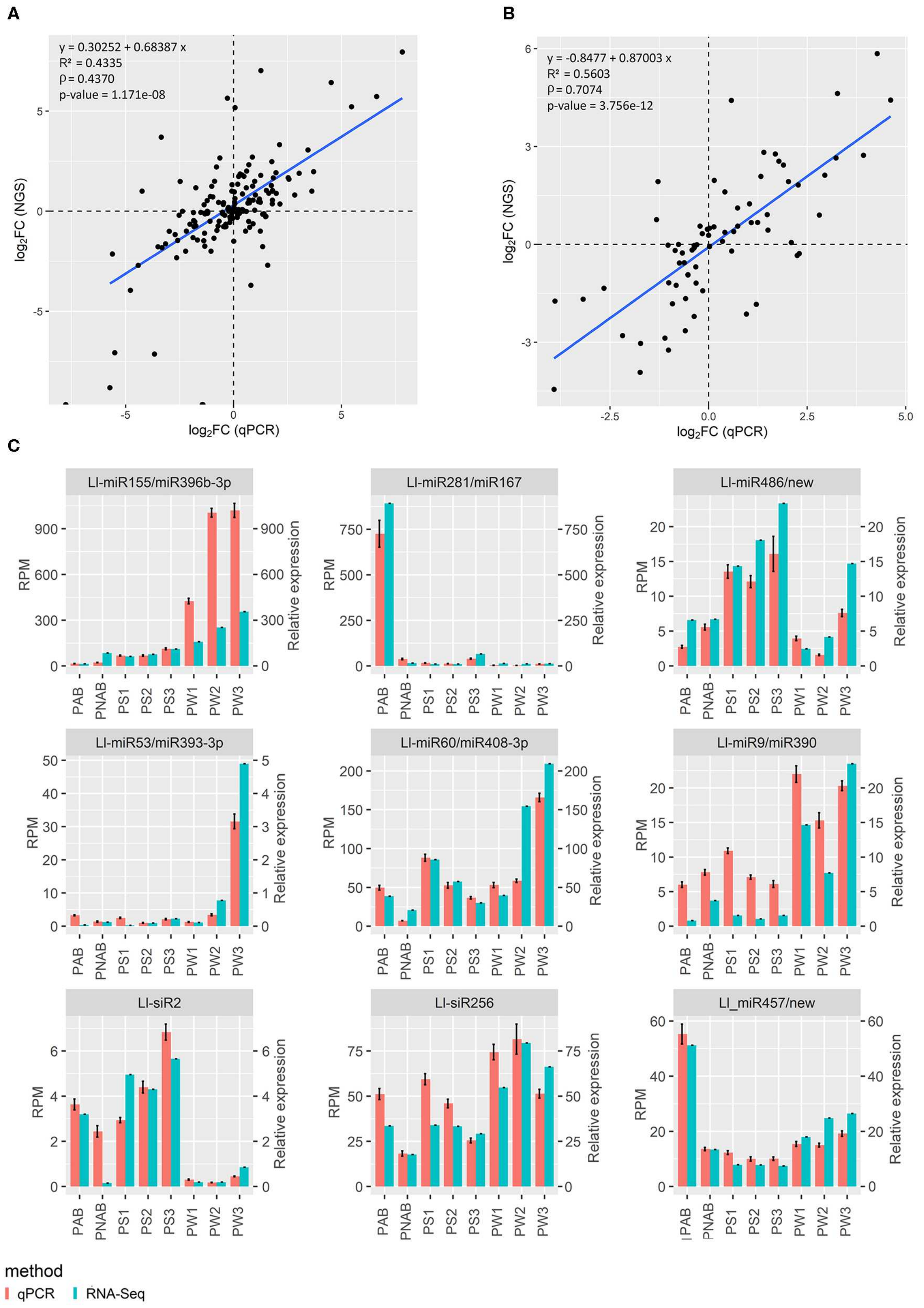


FIGURE 3 | Relationship between next-generation sequencing and qPCR results. **(A)** Log$_2$ fold change of gene expression assessed using NGS plotted against log$_2$ fold change of gene expression assessed using qPCR. **(B)** The same plot for sRNA data. R^2 is coefficient of determination, ρ is Spearman's rank correlation coefficient. **(C)** Graphs showing the similar trend in expression levels of miRNAs and siRNAs assessed with NGS and qPCR.

following parameters: *E*-value Threshold: $1e^{-10}$ and Max Number of Alignments to Report: 100. For respective DCL sequences, different lists of results were acquired (**Supplementary Tables 3–6** for DCL1, 2, 3, and 4, respectively). It is noteworthy that our database allows the users to download BLASTn results. The most homologous sequences were found for *DCL2* and the least were found for *DCL1* and *4*, and in each case, it was found in both tissues, flowers and pods. The pages for individual transcripts contain detailed information, such as ID, sequence, annotation, and expression in different organs, which can be viewed and downloaded, and the "details" link redirects the user to amino acid sequence and additional information. We employed two ways of obtaining the data: manual, where we enter the site of each identified transcript one by one and download all available information, or more automated, where we use Python script to search the files downloaded from the main page with the lists of homologous transcripts from BLASTn. Both ways gave the same results.

All sequences identified by BLASTn are annotated as *DCL* genes (**Supplementary Tables 3–6**). Only the transcripts with complete ORFs and coding the same protein in both tissues, flowers and pods, were used for further analysis. The only exception here is *DCL2*, where the only transcript in flowers with complete ORF has no equivalent in the set of transcripts in pods, which are much shorter.

In *A. thaliana*, four members of DICER-like family are responsible for sncRNA biogenesis of different lengths (Gasciolli et al., 2005). DCL1 is involved mainly in biogenesis of 21 nt miRNA by cutting out miRNA/miRNA* duplexes from imperfect fold back stem-loop structures within pri-miRNA and pre-miRNA precursors (Liu et al., 2005; Song et al., 2011). DCL2, DCL3, and DCL4 are responsible for generating siRNA from dsRNA derived from exogenous elements, natural antisense genes, transcripts of *TAS* genes, and probably *PHAS* genes, or repeated heterochromatic regions (Gasciolli et al., 2005; Henderson et al., 2006; Rajagopalan et al., 2006; Chen et al., 2010). Despite the differences, functions of all DCL enzymes partly overlap (Gasciolli et al., 2005).

In many plant species *DCL* genes are duplicated. For example, rice genome encodes two isoforms of *DCL2* and two of *DCL3* (Kapoor et al., 2008). Similar situation occurs in legumes. In soybean, seven *DCL* genes were described, two isoforms of *DCL1*, *DCL2*, and *DCL4* each and one of *DCL3* (Curtin et al., 2012; Liu et al., 2014b). In *M. truncatula* genome, six *DCL* genes encode members of all four types of DCL (DCL1–4), including three isoforms of DCL2 (Tworak et al., 2016). In *L. japonicus*, five subsequent *LjDCL* genes were identified: *LjDCL1*, *LjDCL2a* and *LjDCL2b*, *LjDCL3*, and *LjDCL4* (Bustos-Sanmamed et al., 2013). Within narrow-leaved lupine genome, seven *DCL* genes were found, including three homologs of *DCL2*, two of *DCL3*, and one of *DCL1* and *DCL4* (DeBoer et al., 2019).

In this work, we have demonstrated that members of all *DCL* families are present and expressed in generative organs in yellow lupine. Our results suggest the presence of more than one gene that codes for each type of DCL protein, similar to other *Fabaceae* (**Figure 4**). On the basis of both nucleotide and amino acid sequences, we were able to distinguish two isoforms

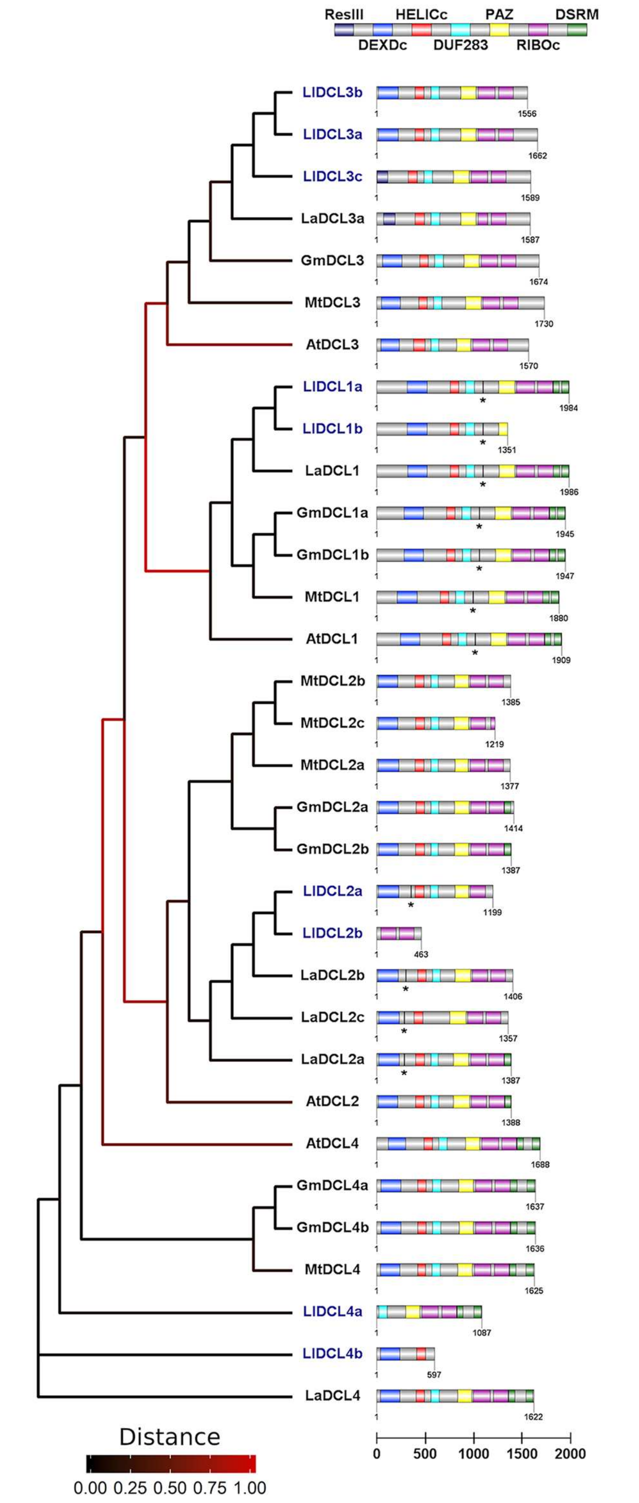


FIGURE 4 | Phylogenetic tree and domain structure of members of DCL families identified in *Lupinus luteus* (Ll), *Arabidopsis thaliana* (At), *Medicago truncatula* (Mt), *Glycine max* (Gm), and *Lupinus angustifolius* (La). Asterisks indicate predicted miRNA target sites. Lupine sequences were translated using SeqIO biopython package, the phylogenetic tree was created using *Phylogeny.fr* interface, and the domain architecture was drawn using DOG v 2.0 software.

of *DCL1*, *DCL2*, and *DCL4*, and three of *DCL3*, all of which show greatest similarity to the corresponding homologs *Lupinus angustifolius* (**Figure 4**). Please note that these results do not show the total number of *DCL* genes in *L. luteus*; they only show which of them are expressed in generative organs. In order to reveal the complete landscape of *DCLs* in yellow lupine, we need to analyze the sequence of its genome, which is still unavailable (Iqbal et al., 2020).

In plants, members of DCL protein families contain six conserved domains: N-terminal helicase domain (built with DEXD/H-box and helicase-C subdomains), followed by DUF283 (domain of unknown function, also known as Dicer-dimer or Dicer dimerization domain), PAZ (Piwi-Argonaute-Zwille), tandemly arranged two RNase III domains, and up to two C-terminal dsRBD (dsRNA binding) domains (Carmell and Hannon, 2004; Margis et al., 2006; Murphy et al., 2008). The main catalytic activity is demonstrated by two RNase III domains, which cleave dsRNA substrates and form short RNA duplexes. PAZ and helicase domains are known to play a role in proper docking of sncRNA precursor within DCL protein (MacRae et al., 2006; Gu et al., 2012), and helicase additionally enables processing of longer substrates (Cenik et al., 2011; Welker et al., 2011). The dsRBD domain is thought to be involved in the recognition and processing of RNA substrates as well as in interactions with other elements of sncRNA biogenesis pathway (Hiraguri et al., 2005; Eamens et al., 2009).

The presence of all of the abovementioned domains in DCL1 is highly conserved across plants including legumes, which proves that it plays the most important role in sncRNA biogenesis (Gasciolli et al., 2005; Parent et al., 2012). The domainal organization of other DCL proteins is more varied in different plant species, which is probably related to their overlapping functions and the resulting increased tolerance to aberrations.

In our analyses, only the *DCL1* transcript encodes protein sequence containing all possible DCL domains (**Figure 4**). In the other cases, the putative proteins are truncated and lack either C- or N-terminal domains. Regarding DCL2 and DCL3, the C-terminal truncation seems to be evolutionarily conserved, and these proteins contain at most one single copy of the DSRM domain, and in the case of MtDCL2c and LlDCL2b—they also lack one of the RIBOc domains. In the set of studied DCL2 homologs, only LlDCL2b is truncated at the N-terminus, and as it lacks all the domains except for two RIBOc, it is probably a pseudogene. In LlDCL3c and LaDCL3a, the N-terminal domain DEXDc is narrowed to sequence encoding the res subunit of type III restrictase (ResIII). It is striking that if we merged LlDCL2a and b, as well as LlDCL4b and a, we would obtain a full-length protein. This fact suggests that some mutations in *DCL2* and *4* homologs in yellow lupine might have occurred, leading to the emergence of non-functional or partly functional proteins. Thus, problems with identifying full-length ORF sequences of *DCL2* in pods may arise because of similar reasons: mutations that change the reading frame or influence alternative splicing, or other unknown causes. It is difficult to determine the physiological effects of the presence of truncated DCL proteins in yellow lupine; however, considering that we have identified a number of siRNAs, e.g., tasiR-ARF (Glazinska et al., 2019), the biogenesis of siRNA in this plant is unaffected. Perhaps the truncated proteins form complexes with more complete DCLs, which enables their participation in sncRNA biogenesis. It would be interesting to examine the exact function of these DCLs.

In addition to analyzing the putative amino acid sequence of DCLs, we have also explored nucleotide sequences of transcripts, which encode these proteins. The mRNA and CDS sequences deposited in the database enable the identification of non-translated sequences, e.g., 5′UTR, which often contain regulatory sequences providing premises for speculation on the possible factors affecting expression of the studied genes. We performed sequence analysis of 5′UTR regions of genes encoding DCL1, 3, and 4 (**Supplementary Table 7**), except for *DCL2*—because of its shortness (∼80 nt). In the case of *DCL4*, the mRNA sequence upstream to the origin of the identified ORF is as long as 1,726 nt, which is an extremely long 5′UTR. This may be associated with the N-terminal truncation of DCL4 protein, probably caused by a mutation that either turned off the original start codon and switched the start of translation downstream to the next ATG, or turned on the stop codon between original start codon and the next ATG. Further analyses support the latter hypothesis, as in FTRINITY_DN57273_c0_g1_i5 transcript; for example, there is an additional ATG at 173 nt located farther than in *A. thaliana*, which may encode a protein containing missing N-terminal domains, but it is stopped prematurely. Moreover, these two ORFs are placed in different reading frames. In the LuluDB, only the longest CDS identified on a given mRNA was deposited. For 5′UTRs of *DCL1* and *3*, they are 604 and 201 nt long, respectively.

We have analyzed selected 5′UTRs by querying PlantCare, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences (Lescot, 2002). All UTR sequences contain typical promoter elements, namely, CAAT (common *cis*-acting element in promoter and enhancer regions) and TATA box (core promoter element around −30 of transcription start) as well as elements related to the response to plant hormones (e.g., IAA, SA, or ABA) or stress factors and light (**Supplementary Table 7**). This demonstrates the complex regulation of expression of these transcripts.

Figure 5 shows the expression of all LuluDB BLASTn hits containing complete ORFs. The overwhelming majority of them show low or even extremely low expression in most of the tested samples. One exception is mRNA encoding DCL1a in both flowers and pods, which exhibits the highest level of expression in stage 3 of flower development, and in the oldest walls and seeds of pods. Other exceptions include single transcripts encoding DCL2 and 3, which accumulate the most in yellow lupine pods. Interestingly, shortened transcripts (containing partial ORFs) of *DCL2* and *DCL4* have higher accumulation (**Supplementary Tables 4**, **6**). It would be extremely interesting to investigate the cause. Our hypothesis, which says that they are products of cutting longer transcripts, requires further analysis.

In soybean, *DCL2* expression is regulated by miR1515 (Li et al., 2010), and in *Medicago* by miR1507 (Zhai et al., 2011). Our analysis shows that in yellow lupine, none of these miRNAs targets any of DCL2-encoding transcripts: miR1515 is missing from our libraries, as well as the target transcripts of miRNA annotated as miR1507 (ID88) code for members of the Disease resistance protein RGA family. Exploring the data deposited in LuluDB can lead to further research opportunities. Our short case study on *DCL* genes, for example, evoked several questions and presented new exciting challenges, including an analysis of changes in the expression of identified genes under the influence of factors predicted by 5′UTRs analysis, or complete characteristics of genes encoding DCLs as soon as good quality genome of *L. luteus* is released.

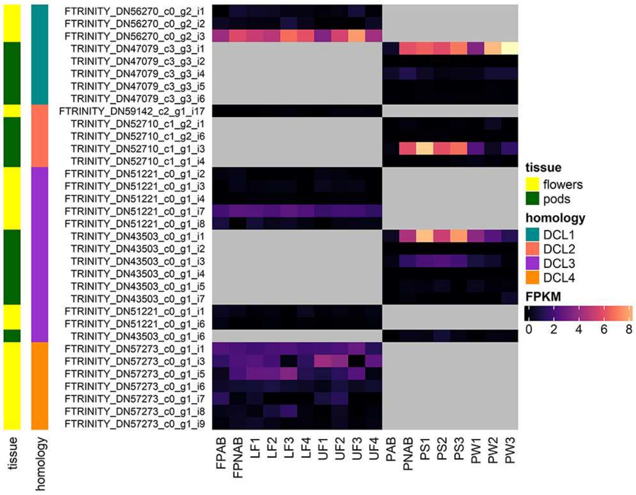


FIGURE 5 | Heatmap presenting expression of RNAs coding for DCLs identified in yellow lupine, created using the "ComplexHeatmap" R package.

Analysis of miRNAs That Target *DCL1* and *DCL2* in Yellow Lupine

Literature data contain evidence that miR162 regulates the *DCL1* expression in other plants (Xie et al., 2003; Liu et al., 2014a; Szajko et al., 2019), whereas in yellow lupine, we have identified a new regulator of *DCL2* (Glazinska et al., 2019), a novel miRNA named "Ll-miRn30," deposited in the database under ID486.

After typing the phrase "miR162" in the search bar of the browse/miRNA section, we are presented with a list of eight lupine miRNAs annotated as miR162, which means that they are identical to miR162s from other plant species deposited in miRBase (**Figure 6**). This indicates that sequences of lupine miR162 are evolutionarily conserved; miRNA of ID239 is most conserved and ubiquitous, as it is annotated to the biggest number of miRNA sequences from miRBase (Kozomara et al., 2019).

L. luteus miR162s are annotated to both miR162-3′ and miR162-5′. The detailed analysis showed that five of them (ID238, 239, 240, 242, and ID417) have expression level higher than 0.2 RPM in most of sequenced small RNA libraries (**Supplementary Table 8**). All five of them are −3p sequences, which are considered as main forms of biologically active miR162 (Kozomara et al., 2019), which indicates that this form is also crucial in yellow lupine. All of them exhibit a similar accumulation profile, with the highest expression in pod walls (PW3), flower pedicels (FPNAB and FPAB), and the lowest in the youngest abscissing and non-abscissing pods (PAB and PNAB) (**Figure 6**). The miR162 of ID239 not only has the greatest number of annotations but also displays the highest expression.

Analyses of the data for miR162 present in LuluDB indicate that this miRNA can also regulate the expression of *DCL1* in yellow lupine (**Supplementary Table 9**). All five miRNAs have a long list of targets identified by both degradome and

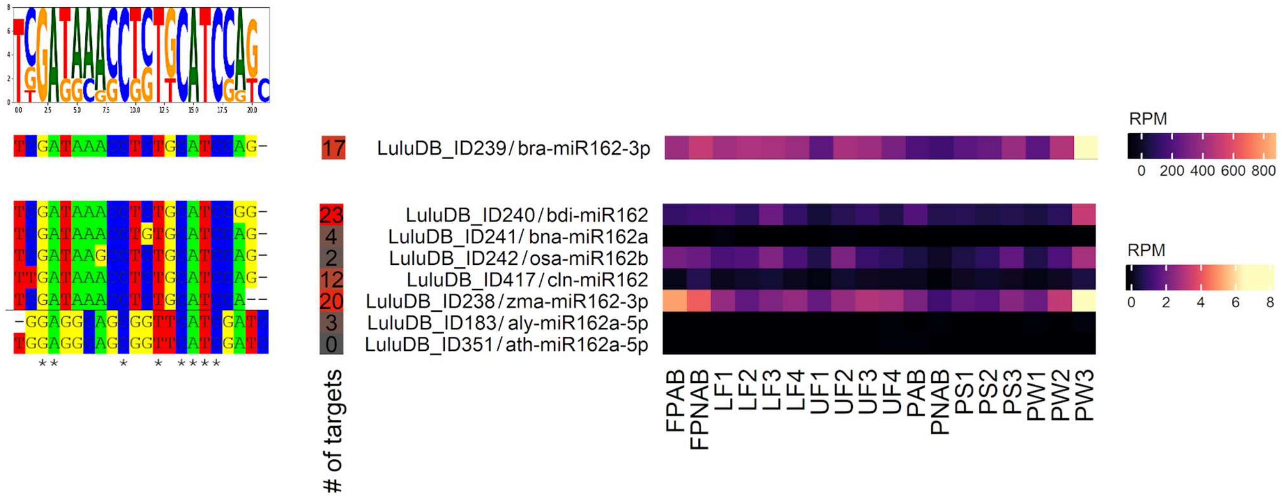


FIGURE 6 | Homologs of miR162 identified in yellow lupine. Left: aligned miRNA sequences with the sequence logo on the top. Right: Heatmap presenting expression of homologs of miR162 identified in yellow lupine, created using the "ComplexHeatmap" R package.

psRNATarget analysis. In each case, we have information about miRNA binding sites, which are located near the center of transcripts in the region devoid of any functional protein domains (**Figure 4**), similarly to mRNA coding DCL1 in other plant species (Xie et al., 2003; Shao et al., 2015).

It is noteworthy that the download feature is very useful for both analysis and presentation of the data. The downloaded files include information about miRNA ID, its sequence, the list of target genes, and expression of miRNA in different organs. Redirection links to pages for individual targets show that majority of them were identified in degradomes and that they are in fact transcripts annotated as *DCL1*. They are present in both flowers and pods (**Supplementary Table 9**). Most of them exhibit low level of accumulation, accompanied by high expression of miR162 (ID239), which suggests a negative correlation and tight regulation of *DCL1* mRNA levels by its degradation, or localization of the transcript limited to specific type of cells. Interestingly, sequence miR162-5p (ID183) also has degradome-identified targets in yellow lupine annotated as putative clathrin assembly protein (**Supplementary Table 9**); however, given the fact that expression levels of this miRNA are very low, its regulatory potential is highly unlikely.

Regarding the already mentioned novel miRNA ID486, which is most likely responsible for regulation of *DCL2* in yellow lupine, there is a long list of targets found by degradome analysis in both flowers and pods in the database (**Table 3**). Most of them are annotated as *DCL2*; however, some of them encode homologs of Nucleolar protein 12 (NOL12). NOL12 is a protein described, e.g., in humans as a multifunctional protein (Scott et al., 2017): with its ability to bind rRNA, it is required for efficient separation of large and small subunit precursors, and by binding with DNA repair proteins, it is essential for genome integrity. Interestingly, the potential target site for miRNA ID486 is also present within *DCL2* sequence of narrow-leaved lupine (XM_019584571.1) (**Figure 7A**), which can indicate that this regulator is characteristic for this closely related lupine species. The expression of miRNA ID486 is the highest in LF1 and decreases during flower development, contrary to upper flowers that exhibit the highest level of expression at stage 3 (UF3). In the case of pods, it is strongly accumulated

TABLE 3 | List of target genes for novel miR486 from *Lupinus luteus*.

LuluDB transcript	Tissue	Target sequence	Target start position	Target end position	Regulation type	Transcript annotation
Ll_transcript_256739	Flowers	GCAGAGTCTGCACACAAACGAA	903	924	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256747	Flowers	GCAGAGTCTGCACACAAGCGAA	903	924	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256729	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256731	Flowers	GCAGAGTCTGCACACAAACGAA	903	924	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256736	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479081	Pods	GCAGAGTCTGCACACAAACGAA	1131	1152	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479083	Pods	GCAGAGTCTGCACACAAACGAA	1491	1512	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479084	Pods	GCAGAGTCTGCACACAAACGAA	1020	1041	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479065	Pods	GCAGAGTCTGCACACAAACGAA	1131	1152	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479068	Pods	GCAGAGTCTGCACACAAACGAA	919	940	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256739	Flowers	GCAGAGTCTGCACACAAACGAA	903	924	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256726	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256727	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256729	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256734	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_256736	Flowers	GCAGAGTCTGCACACAAACGAA	1115	1136	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479081	Pods	GCAGAGTCTGCACACAAACGAA	1131	1152	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479083	Pods	GCAGAGTCTGCACACAAACGAA	1491	1512	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479084	Pods	GCAGAGTCTGCACACAAACGAA	1020	1041	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479065	Pods	GCAGAGTCTGCACACAAACGAA	1131	1152	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_479068	Pods	GCAGAGTCTGCACACAAACGAA	919	940	Cleavage	Endoribonuclease Dicer homolog 2
Ll_transcript_219270	Flowers	TCAGGGTCTGCGCGCAAACAAA	2411	2432	Cleavage	Nucleolar protein 12
Ll_transcript_219271	Flowers	TCAGGGTCTGCGCGCAAACAAA	1000	1021	Cleavage	Nucleolar protein 12
Ll_transcript_219276	Flowers	TCAGGGTCTGCGCGCAAACAAA	657	678	Cleavage	Nucleolar protein 12
Ll_transcript_219278	Flowers	TCAGGGTCTGCGCGCAAACAAA	1089	1110	Cleavage	Nucleolar protein 12
Ll_transcript_386361	Pods	TCAGGGTCTGCGCGCAAACAAA	2294	2315	Cleavage	Nucleolar protein 12
Ll_transcript_386366	Pods	TCAGGGTCTGCGCGCAAACAAA	2412	2433	Cleavage	Nucleolar protein 12
Ll_transcript_386378	Pods	TCAGGGTCTGCGCGCAAACAAA	1255	1276	Cleavage	Nucleolar protein 12
Ll_transcript_386378	Pods	TCAGGGTCTGCGCGCAAACAAA	1255	1276	Cleavage	Nucleolar protein 12

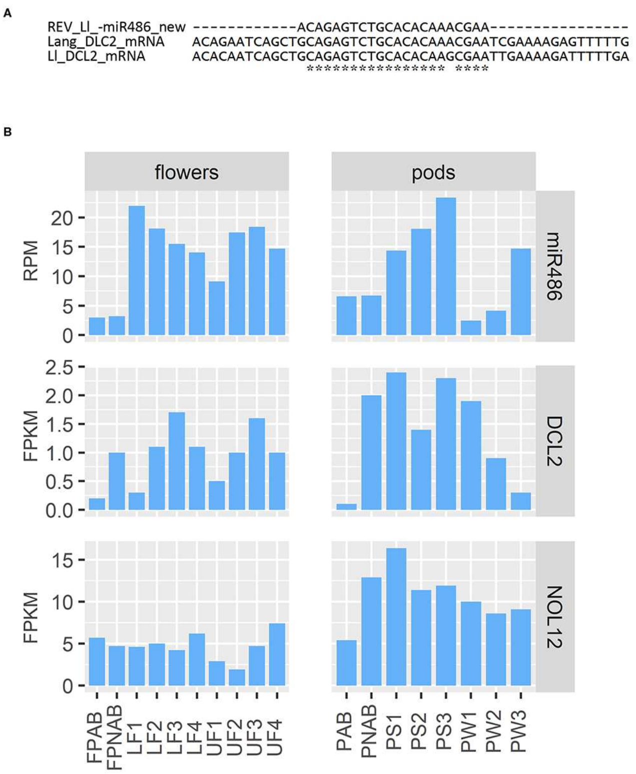


FIGURE 7 | Analysis of miR486. **(A)** Comparison of target sequence for miR486 in DCL2-encoding transcript in yellow and narrow-leaved lupines. **(B)** Expression of miR486 and its selected target genes.

in the oldest seeds (PS3) (**Figure 7B**). Expression of its target genes is the highest in flowers at stages 3 and 4 independently from their position on inflorescence, as well as in young seeds (PS1), so it is exhibiting a reverse tendency in comparison to miRNA (**Figure 7B**).

Homologs of both miR162 as well as new miR486 do not show differential expression in pedicels of abscissing and non-abscissing flowers similarly to abscissing and non-abscissing pods, which can indicate that they are not directly linked to the generative organ abscission process in yellow lupine. However, changes in their accumulation during the development suggest that they regulate the sRNA biogenesis depending on the stage of development in both flowers and pods.

Other Examples and Suggestions for the Use of Data Contained in the LuluDB

The data present in the database was already used to identify new mechanisms for regulating gene expression by sRNA in yellow lupine; e.g., we described the involvement of sRNAs in *L. luteus* flower development including new miRNAs and the new siRNA (siR240), which, together with the conserved tasiR-ARF, may trigger cleavage of the *TAS3* transcript (Glazinska et al., 2019). Data present in the database were also used to perform analyses published by Shi and coworkers (Shi et al., 2019), where the authors identified yellow lupine *IDA* (*INFLORESCENCE DEFICIENT IN ABSCISSION*), *IDL* (*IDA-LIKE*), and *HSL* (*HAESA*) transcripts, and compared their expression with their homologs in other plant species. The authors proved that the lupine gene most similar to *AtIDA* (*LlIDA*) plays the most important role in lupine AZs compared to the *LlIDL* genes. This work indicates that the mechanism of the generative organ abscission in lupine and *A. thaliana* has common features. These papers were published before the LuluDB was made public and we believe that publication of an article on our database will encourage more researchers to use it for their own purposes.

To date, the LuluDB is profiled to provide information about the regulation of transcripts by miRNA and siRNA, confirmed by degradome analysis. However, it also contains a category of "lncRNA," which was not explored by our group, and gives the opportunity to perform preliminary *in silico* analyses of this kind of regulatory factors. Additionally, a significant amount of transcripts are not annotated to any database. The algorithms for protein–protein and protein–ligand docking are constantly being improved, giving the opportunity to specify the function of chosen protein basing solely on its sequence. Today, guessing that the function of all the unannotated protein-coding transcripts would be impossible—this method is time- and memory-consuming. However, in the near future, computers are going to be better equipped and algorithms are going to be faster, so probably this kind of analyses will be performed.

CONCLUSION

The presented LuluDB database has been equipped with user-friendly and intuitive tools for searching and investigating our NGS data, including more advanced bioinformatics. **Figure 8** depicts possible approaches to explore the database and navigate its components, as well as their interlinked sections. The presented case study visualizes how the use of LuluDB database improves the process of analysis of both miRNAs and their target genes involved in regulation of growth or abscission of generative organs in yellow lupine, and allows for identification of homologs of selected genes and evolutionary analyses. The database can be used as a starting point for different types of research concerning protein-coding RNAs and ncRNAs not only for yellow lupine but also for other plant species.

MATERIALS AND METHODS

Plant Material

Yellow lupine plants used for RNA extraction were cultivated in Nicolaus Copernicus University's experimental field in Piwnice near Torun (Poland, 53°05′42.0″N 18°33′24.6″E), as described in detail in Glazinska et al. (2019). The flowers were separated based on their developmental stage and position on raceme, as described previously in Glazinska et al. (2019). Additionally, flower pedicles from abscissing (FPAB) and

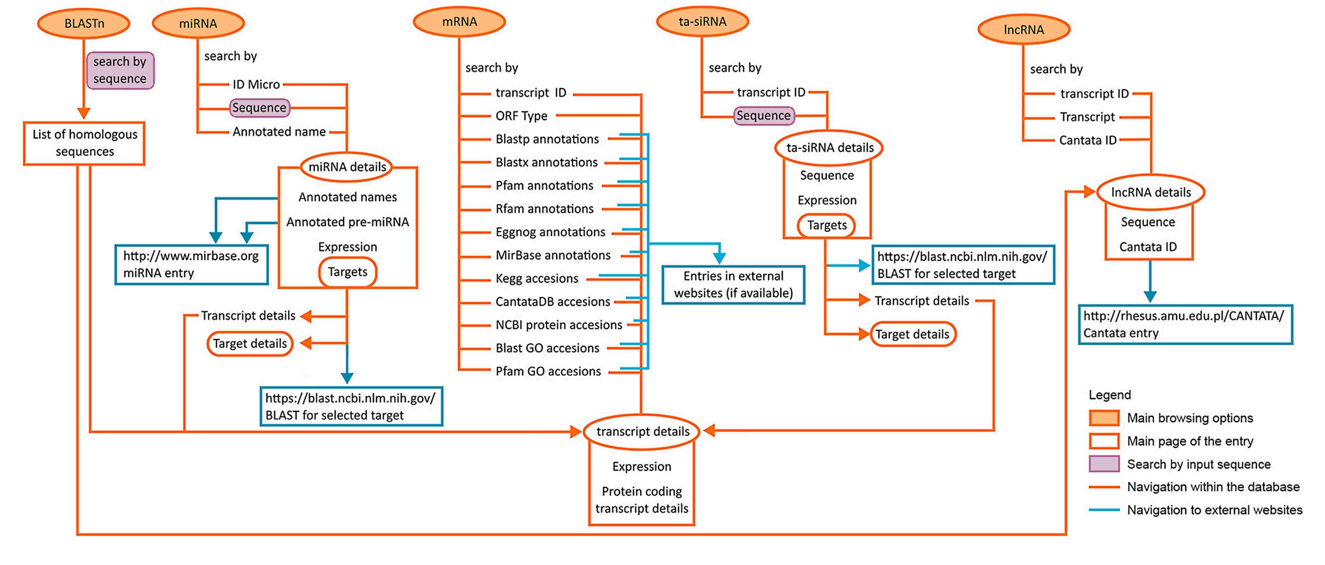


FIGURE 8 | Diagram depicting the possible ways to access and explore the database.

non-abscissing flowers (FPNAB) were collected. Then, pods in different stages of development were also harvested and separated into pod walls (PW) and seeds (PS). Only the pods from the lowest whorl were collected, based on the fact that most of the pods set at higher whorls undergo abscission at early stage of development (van Steveninck, 1957). In addition, entire small abscissing pods (PAB) as well as non-abscissing pods of the same length (PNAB) were collected as a whole.

RNA Isolation, Library Construction, and NGS of Small RNA, Transcriptome, and Degradome

RNA isolation from all of the collected samples was carried out using miRNeasy Mini Kit (Qiagen, Venlo, the Netherlands) with on-column DNA digestion with the RNase-Free DNase Set (Qiagen, Venlo, the Netherlands), as described in detail in Glazinska et al. (2019). After passing all quality and quantity checks (as described in Glazinska et al., 2019), total RNA was used for preparation of small RNA libraries using NEBNext Multiplex Small RNA Library Prep kit for Illumina (New England Biolabs, Ipswich, MA, USA) and subsequently sequenced on the HiSeq4000 platform (Illumina, San Diego, CA, USA) in the 50 single-end mode.

Similarly isolated total RNA was used to create transcript libraries using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) and sequenced on the HiSeq4000 platform in the 100 paired-end mode as described in detail in Glazinska et al. (2019).

Degradomes were obtained using total RNA pooled from samples UF3/LF3 and PW3/PS3 to meet the amount of material required for sequencing. The protocol for degradome library preparation and detailed information can be found in Glazinska et al. (2019). The small RNA libraries were created from two biological replicates, transcriptomes from three pooled replicates, and degradomes from pooled samples of LF3–UF3 and PS3–PW3; thus, in total, we have sequenced 32 small RNA, 16 transcriptome, and 2 degradome libraries.

RNA-Seq: *de novo* Transcriptome Assembly and Transcript Expression Estimation

The *de novo* transcriptome assembly was performed on RNA-Seq data using Trinity v 2.4.0 (https://github.com/trinityrnaseq/trinityrnaseq/releases) with default settings as described in Glazinska et al. (2017). The transcriptome assembly for flowers and pods was carried out separately. Therefore, to make this division clear, identified transcripts from flowers have an additional letter "F" as a prefix (e.g., FTRINITY_DN53848_c2_g1_i5). Estimation of expression level on both the unigene and isoform level was reported in FPKM and was done using RSEM (Haas et al., 2013) as described in Glazinska et al. (2017).

RNA-Seq: Annotating the Transcriptomes

Annotation of transcriptomes was performed with Trinotate (v 3.0.2). BLASTX with "max_target_seqs 1" option was used to identify the sequence similarity between lupine transcripts and proteins annotated in Swiss-Prot, a non-redundant and manually curated dataset from the UniProt database. Open reading frames were predicted with TransDecoder (v 5.0.1) (Haas et al., 2013) in order to scan inferred protein sequences against Swiss-Prot using BLASTP (with "-max_target_seqs 1" option). hmmscan (hmmer.org) with default settings was used to identify protein domains based on the similarity to Pfam database records (http://pfam.xfam.org/). All of the previously mentioned results were loaded into an SQLite database built by Trinotate and used to generate the final report. Another approach for annotating the transcriptomes was also applied using sole BLAST searches against public databases, which included BLASTN against RFAM (no *E*-value threshold), miRBase (*E*-value threshold of $1e^{-5}$), and *G. max* lncRNAs from CantataDB (Szcześniak et al., 2016) (*E*-value threshold of $1e^{-5}$). Additionally, the BLASTX search against the NCBI protein database (nr) for *Fabaceae* (no *E*-value threshold) was performed. The results were then parsed to obtain the best hits per transcript based on the alignment score value.

Identification of Small ncRNAs and Their Target Genes

To identify phylogenetically conserved mature miRNAs with sequences and lengths identical to known plant miRNAs, we searched miRBase for similarity at the mature miRNA level. Short reads from RNA-Seq were compared against mature miRNAs from miRBase (Kozomara et al., 2019). The comparison was performed with Bowtie (Langmead, 2010), allowing for no mismatches. To predict potential novel miRNAs, we applied ShortStack (Axtell, 2013b) with default settings as described in Glazinska et al. (2019). ShortStack (Axtell, 2013b) was also used to identify small RNAs that were being cut in phase from longer precursors (phased siRNAs) (Glazinska et al., 2019). The top 200 candidates were selected from each sample, based on the phased score value provided by ShortStack. Finally, lists of such sRNAs from all samples were merged into a single dataset of non-redundant phased siRNAs. The expression of sncRNA was presented in RPM units. Each discovered miRNA received an identification ID number. MiRNAs identified using ShortStack and not showing sequence similarity with miRBase were given numbers from 457 up and have the annotation "none" in the database. Target genes for identified small RNAs were estimated based on degradome data, or using psRNATarget tool as described in Glazinska et al. (2019). The miRNA targets were searched among assembled transcriptomes, separately for flowers and for pods. The same analysis was done for siRNAs.

Expression Analysis With RT-qPCR

MiRNA and siRNA expression was analyzed using the Stem Loop RT-qPCR technique according to Glazinska et al. (2019). Expression of protein coding transcripts was measured as in Glazinska et al. (2017). Each experiment consisted of three biological and technical replicates. The relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method, and the data were normalized to the CT values for the *LlActin4* reference gene (according to Glazinska et al., 2017). All primer sequences are listed in **Supplementary Table 10**. To assess the linearity of relationship and correlation strength between sRNA-Seq or RNA-Seq and qPCR data, we have first log-transformed the data and calculated R^2 and Spearman's rank correlation coefficient (ρ), respectively, using R packages: dplyr, ggpubr, and Hmisc.

Data Submission to Sequence Read Archive (NCBI)

The RNA-Seq data and small RNA-Seq data have been uploaded to SRA database and are available under BioProject ID PRJNA419564 and Submission ID SUB3230840.

Database Implementation and Testing

LuluDB was developed using Hypertext Markup Language (HTML), Sassy Cascaded Style Sheets (SCSS), Cascading Style Sheets (CSS), PHP 5.6, Yii 2.0 PHP framework (https://www.yiiframework.com/), MySQL 5.5, JavaScript, jQuery 3.2.1 (https://jquery.com/), MSAViewer (Yachdav et al., 2016), and Bootstrap 3.3.7 framework (https://getbootstrap.com/). NCBI BLAST+ 2.7.1 (Camacho et al., 2009) was used as a local alignment search tool. Database can be run and has been tested on most currently widely used web browsers regardless of operating system, including Firefox Web Browser, Safari, Google Chrome, and Opera. Responsive web design was applied to ensure that the database will be properly displayed on mobile devices.

DATA AVAILABILITY STATEMENT

All datasets analyzed for this study are included either in the article/**Supplementary Material** or in the database at http://luluseqdb.umk.pl/basic/web/index.php. The RNA-Seq data and small RNA-Seq data have been uploaded to SRA database and are available under BioProject ID PRJNA419564 and Submission ID SUB3230840.

AUTHOR CONTRIBUTIONS

PG: conceptualization, funding acquisition, supervision, and writing—review and editing. PG, MW, and JK: data curation. PG, MK, and WG: investigation, visualization, and writing—original draft. JK and MW: software.

FUNDING

This research was funded by The National Science Centre SONATA grant No. 2015/19/D/NZ9/03601 and by the Program Supported by Resolution of the Council of Ministers (RM-111-222-15) in association with the Institute of Plant Genetics, Polish Academy of Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00455/full#supplementary-material

Supplementary Figure 1 | Alignment of sequences of corresponding DCL1-coding transcripts expressed in flowers and pods.

Supplementary Figure 2 | Screenshot of LuluDB page concerning: **(A)** phased-siRNA, **(B)** long non-coding RNA.

Supplementary Figure 3 | Screenshot of LuluDB page concerning protein-coding RNA sequence.

Supplementary Figure 4 | Juxtaposition of NGS and qPCR expression levels of eight transcripts used for validation. Homologs of the same transcript found in flowers and pods are shown separately.

Supplementary Table 1 | Details of data deposition in NCBI SRA.

Supplementary Table 2 | A list of DCL sequences from *Arabidopsis thaliana* and selected *Fabaceae* plants used for identification of homologs in *L. luteus* and in phylogenetic analyses.

Supplementary Table 3 | Results of LuluDB search by built-in BLASTn using *A. thaliana DCL1* CDS sequence as query.

Supplementary Table 4 | Results of LuluDB search by built-in BLASTn using *A. thaliana DCL2* CDS sequence as query.

Supplementary Table 5 | Results of LuluDB search by built-in BLASTn using *A. thaliana DCL3* CDS sequence as query.

Supplementary Table 6 | Results of LuluDB search by built-in BLASTn using *A. thaliana DCL4* CDS sequence as query.

Supplementary Table 7 | List of regulatory sequences identified within 5′UTRs of *L. luteus* mRNAs coding for DCL1, 3 and 4 using PlantCare search.

Supplementary Table 8 | A detailed list of miRNAs in LuluDB annotated as miR162.

Supplementary Table 9 | List of target transcripts for members of *MIR162* family deposited in LuluDB.

Supplementary Table 10 | List of primers and UPL probes used for RT-qPCR reaction.

REFERENCES

Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25, 130–131. doi: 10.1093/bioinformatics/btn604

Axtell, M. J. (2013a). Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* 64, 137–159. doi: 10.1146/annurev-arplant-050312-120043

Axtell, M. J. (2013b). ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 19, 740–751. doi: 10.1261/rna.035279.112

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi: 10.1016/s0092-8674(04)00045-5

Bustos-Sanmamed, P., Bazin, J., Hartmann, C., Crespi, M., and Lelandais-Brière, C. (2013). Small RNA pathways and diversity in model legumes: lessons from genomics. *Front. Plant Sci.* 4:236. doi: 10.3389/fpls.2013.00236

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Carmell, M. A., and Hannon, G. J. (2004). RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.* 11, 214–218. doi: 10.1038/nsmb729

Cenik, E. S., Fukunaga, R., Lu, G., Dutcher, R., Wang, Y., Tanaka Hall, T. M., et al. (2011). Phosphate and R2D2 restrict the substrate specificity of Dicer-2, an ATP-driven ribonuclease. *Mol. Cell* 42, 172–184. doi: 10.1016/j.molcel.2011.03.002

Chen, H. M., Chen, L. T., Patel, K., Li, Y. H., Baulcombe, D. C., and Wu, S. H. (2010). 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15269–15274. doi: 10.1073/pnas.1001738107

Curtin, S. J., Kantar, M. B., Yoon, H. W., Whaley, A. M., Schlueter, J. A., and Stupar, R. M. (2012). Co-expression of soybean Dicer-like genes in response to stress and development. *Funct. Integr. Genomics* 12, 671–682. doi: 10.1007/s10142-012-0278-z

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46, W49–W54. doi: 10.1093/nar/gky316

DeBoer, K., Melser, S., Sperschneider, J., Kamphuis, L. G., Garg, G., Gao, L.-L., et al. (2019). Identification and profiling of narrow-leafed lupin (*Lupinus angustifolius*) microRNAs during seed development. *BMC Genomics* 20:8. doi: 10.1186/s12864-019-5521-8

Eamens, A. L., Smith, N. A., Curtin, S. J., Wang, M. B., and Waterhouse, P. M. (2009). The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA* 15, 2219–2235. doi: 10.1261/rna.1646909

Fei, Q., Xia, R., and Meyers, B. C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell* 25, 2400–2415. doi: 10.1105/tpc.113.114652

Frankowski, K., Wilmowicz, E., Kućko, A., Zienkiewicz, A., Zienkiewicz, K., and Kopcewicz, J. (2015). Molecular cloning of the BLADE-ON-PETIOLE gene and expression analyses during nodule development in Lupinus luteus. *J. Plant Physiol.* 179, 35–39. doi: 10.1016/j.jplph.2015.01.019

Fukudome, A., and Fukuhara, T. (2017). Plant dicer-like proteins: double-stranded RNA-cleaving enzymes for small RNA biogenesis. *J. Plant Res.* 130, 33–44. doi: 10.1007/s10265-016-0877-1

Gasciolli, V., Mallory, A. C., Bartel, D. P., and Vaucheret, H. (2005). Partially redundant functions of arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr. Biol.* 15, 1494–1500. doi: 10.1016/j.cub.2005.07.024

Glazinska, P., Kulasek, M., Glinkowski, W., Wojciechowski, W., and Kosinski, J. (2019). Integrated analysis of small RNA, transcriptome and degradome sequencing provides new insights into floral development and abscission in yellow lupine (*Lupinus luteus* l.). *Int. J. Mol. Sci.* 20:E5122. doi: 10.3390/ijms20205122

Glazinska, P., Wojciechowski, W., Kulasek, M., Glinkowski, W., Marciniak, K., Klajn, N., et al. (2017). De novo transcriptome profiling of flowers, flower pedicels and pods of *Lupinus luteus* (yellow lupine) reveals complex expression changes during organ abscission. *Front. Plant Sci.* 8:641. doi: 10.3389/fpls.2017.00641

Gu, S., Jin, L., Zhang, Y., Huang, Y., Zhang, F., Valdmanis, P. N., et al. (2012). The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing *in vivo*. *Cell* 151, 900–911. doi: 10.1016/j.cell.2012.09.042

Gupta, N., Zahra, S., Singh, A., and Kumar, S. (2018). PVsiRNAdb: A database for plant exclusive virus-derived small interfering RNAs. *Database* 2018, 1–8. doi: 10.1093/database/bay105

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Henderson, I. R., Zhang, X., Lu, C., Johnson, L., Meyers, B. C., Green, P. J., et al. (2006). Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat. Genet.* 38, 721–725. doi: 10.1038/ng1804

Hiraguri, A., Itoh, R., Kondo, N., Nomura, Y., Aizawa, D., Murai, Y., et al. (2005). Specific interactions between Dicer-like proteins and HYL1/DRB-family dsRNA-binding proteins in Arabidopsis thaliana. *Plant Mol. Biol.* 57, 173–188. doi: 10.1007/s11103-004-6853-5

Iqbal, M. M., Erskine, W., Berger, J. D., Udall, J. A., and Nelson, M. N. (2020). "Genomics of yellow lupin (*Lupinus luteus* L.)," in *The Lupin Genome. Compendium of Plant Genomes*, eds K. Singh, L. Kamphuis, and M. Nelson (Cham: Springer), 151–159.

Joshi, T., Wang, J., Zhang, H., Chen, S., Zeng, S., Xu, B., et al. (2017). The evolution of soybean knowledge base (SoyKB). *Methods Mol. Biol.* 1533, 149–159. doi: 10.1007/978-1-4939-6658-5_7

Kapoor, M., Arora, R., Lama, T., Nijhawan, A., Khurana, J. P., Tyagi, A. K., et al. (2008). Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. *BMC Genomics* 9:451. doi: 10.1186/1471-2164-9-451

Kawahara, Y., Oono, Y., Wakimoto, H., Ogata, J., Kanamori, H., Sasaki, H., et al. (2016). TENOR: database for comprehensive mRNA-Seq experiments in rice. *Plant Cell Physiol.* 57:e7. doi: 10.1093/pcp/pcv179

Kim, E., Hwang, S., and Lee, I. (2017). SoyNet: A database of co-functional networks for soybean Glycine max. *Nucleic Acids Res.* 45, D1082–D1089. doi: 10.1093/nar/gkw704

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. doi: 10.1093/nar/gky1141

Kramer, M. F. (2011). Stem-Loop RT-qPCR for miRNAs. *Curr. Protoc. Mol. Biol.* 95, 15.10.1–15.10.15. doi: 10.1002/0471142727.mb1510s95

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics* Chapter 11:Unit 11.7. doi: 10.1002/0471250953.bi1107s32

Lescot, M. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Li, H., Deng, Y., Wu, T., Subramanian, S., and Yu, O. (2010). Misexpression of miR482, miR1512, and miR1515 increases soybean nodulation. *Plant Physiol.* 153, 1759–1770. doi: 10.1104/pp.110.156950

Li, J., Dai, X., Zhuang, Z., and Zhao, P. X. (2016). LegumeIP 2.0–a platform for the study of gene function and genome evolution in legumes. *Nucleic Acids Res.* 44, D1189–D1194. doi: 10.1093/nar/gkv1237

Liu, B., Li, P. C., Li, X., Liu, C. Y., Cao, S. Y., Chu, C. C., et al. (2005). Loss of function of OsDCL1 affects microRNA accumulation and causes developmental defects in rice. *Plant Physiol.* 139, 296–305. doi: 10.1104/pp.105.063420

Liu, H., Jin, T., Liao, R., Wan, L., Xu, B., Zhou, S., et al. (2012). MiRFANs: an integrated database for *Arabidopsis thaliana* microRNA function annotations. *BMC Plant Biol.* 12:68. doi: 10.1186/1471-2229-12-68

Liu, H., Qin, C., Chen, Z., Zuo, T., Yang, X., Zhou, H., et al. (2014a). Identification of miRNAs and their target genes in developing maize ears by combined small RNA and degradome sequencing. *BMC Genomics* 15:25. doi: 10.1186/1471-2164-15-25

Liu, X., Hao, L., Li, D., Zhu, L., and Hu, S. (2015). Long non-coding RNAs and their biological roles in plants. *Genom. Proteom. Bioinform.* 13, 137–147. doi: 10.1016/j.gpb.2015.02.003

Liu, X., Lu, T., Dou, Y., Yu, B., and Zhang, C. (2014b). Identification of RNA silencing components in soybean and sorghum. *BMC Bioinformatics* 15:4. doi: 10.1186/1471-2105-15-4

Lucas, M. M., Stoddard, F. L., Annicchiarico, P., Frías, J., Martínez-Villaluenga, C., Sussmann, D., et al. (2015). The future of lupin as a protein crop in Europe. *Front. Plant Sci.* 6:705. doi: 10.3389/fpls.2015.00705

MacRae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., et al. (2006). Structural basis for double-stranded RNA processing by Dicer. *Science* 311, 195–198. doi: 10.1126/science.1121638

Marchese, F. P., Raimondi, I., and Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* 18:206. doi: 10.1186/s13059-017-1348-2

Margis, R., Fusaro, A. F., Smith, N. A., Curtin, S. J., Watson, J. M., Finnegan, E. J., et al. (2006). The evolution and diversification of Dicers in plants. *FEBS Lett.* 580, 2442–2450. doi: 10.1016/j.febslet.2006.03.072

Murphy, D., Dancis, B., and Brown, J. R. (2008). The evolution of core proteins involved in microRNA biogenesis. *BMC Evol. Biol.* 8:92. doi: 10.1186/1471-2148-8-92

Musco, N., Cutrignelli, M. I., Calabrò, S., Tudisco, R., Infascelli, F., Grazioli, R., et al. (2017). Comparison of nutritional and antinutritional traits among different species (*Lupinus albus* L., *Lupinus luteus* L., *Lupinus angustifolius* L.) and varieties of lupin seeds. *J. Anim. Physiol. Anim. Nutr. (Berl).* 101, 1227–1241. doi: 10.1111/jpn.12643

Ogura, T., Ogihara, J., Sunairi, M., Takeishi, H., Aizawa, T., Olivos-Trujillo, M. R., et al. (2014). Proteomic characterization of seeds from yellow lupin (*Lupinus luteus* L.). *Proteomics* 14, 1543–1546. doi: 10.1002/pmic.201300511

Parent, J.-S., Martínez de Alba, A. E., and Vaucheret, H. (2012). The origin and effect of small RNA signaling in plants. *Front. Plant Sci.* 3:179. doi: 10.3389/fpls.2012.00179

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* 20, 3407–3425. doi: 10.1101/gad.1476406

Robinson, A. J., Tamiru, M., Salby, R., Bolitho, C., Williams, A., Huggard, S., et al. (2018). AgriSeqDB: an online RNA-Seq database for functional studies of agriculturally relevant plant species. *BMC Plant Biol.* 18:200. doi: 10.1186/s12870-018-1406-2

Scott, D. D., Trahan, C., Zindy, P. J., Aguilar, L. C., Delubac, M. Y., Van Nostrand, E. L., et al. (2017). Nol12 is a multifunctional RNA binding protein at the nexus of RNA and DNA metabolism. *Nucleic Acids Res.* 45, 12509–12528. doi: 10.1093/nar/gkx963

Shao, F., Qiu, D., and Lu, S. (2015). Comparative analysis of the Dicer-like gene family reveals loss of miR162 target site in SmDCL1 from *Salvia miltiorrhiza*. *Sci. Rep.* 5:9891. doi: 10.1038/srep09891

Shi, C., Alling, R. M., Hammerstad, M., and Aalen, R. B. (2019). Control of organ abscission and other cell separation processes by evolutionary conserved peptide signaling. *Plants* 8:225. doi: 10.3390/plants8070225

Song, Q.-X., Liu, Y.-F., Hu, X.-Y., Zhang, W.-K., Ma, B., Chen, S.-Y., et al. (2011). Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. *BMC Plant Biol.* 11:5. doi: 10.1186/1471-2229-11-5

Szajko, K., Yin, Z., and Marczewski, W. (2019). Accumulation of miRNA and mRNA targets in potato leaves displaying temperature-dependent responses to Potato Virus Y. *Potato Res.* 62, 379–392. doi: 10.1007/s11540-019-9417-4

Szcześniak, M. W., Bryzghalov, O., Ciomborowska-Basheer, J., and Makałowska, I. (2019). CANTATAdb 2.0: expanding the collection of plant long noncoding RNAs. *Methods Mol. Biol.* 1933, 415–429. doi: 10.1007/978-1-4939-9045-0_26

Szcześniak, M. W., Rosikiewicz, W., and Makałowska, I. (2016). CANTATAdb: A collection of plant long non-coding RNAs. *Plant Cell Physiol.* 57:e8. doi: 10.1093/pcp/pcv201

Tworak, A., Urbanowicz, A., Podkowinski, J., Kurzynska-Kokorniak, A., Koralewska, N., and Figlerowicz, M. (2016). Six Medicago truncatula Dicer-like protein genes are expressed in plant cells and upregulated in nodules. *Plant Cell Rep.* 35, 1043–1052. doi: 10.1007/s00299-016-1936-8

van Steveninck, R. F. (1959). Abscission-accelerators in lupins (*Lupinus luteus* L.). *Nature* 183, 1246–1248. doi: 10.1038/1831246a0

van Steveninck, R. F. M. (1957). Factors affecting the abscission of reproductive organs in yellow lupins (*Lupinus luteus* L.). *J. Exp. Bot.* 8, 373–381. doi: 10.1093/jxb/8.3.373

Varkonyi-Gasic, E., and Hellens, R. P. (2011). Quantitative stem-loop RT-PCR for detection of microRNAs. *Methods Mol. Biol.* 744, 145–157. doi: 10.1007/978-1-61779-123-9_10

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev.* 20, 759–771. doi: 10.1101/gad.1410506

Welker, N. C., Maity, T. S., Ye, X., Aruscavage, P. J., Krauchuk, A. A., Liu, Q., et al. (2011). Dicer's helicase domain discriminates dsRNA termini to promote an altered reaction mode. *Mol. Cell* 41, 589–599. doi: 10.1016/j.molcel.2011.02.005

Xie, Z., Kasschau, K. D., and Carrington, J. C. (2003). Negative feedback regulation of Dicer-Like1 in Arabidopsis by microRNA-guided mRNA degradation. *Curr. Biol.* 13, 784–789. doi: 10.1016/s0960-9822(03)00281-1

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., et al. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* 32, 3501–3503. doi: 10.1093/bioinformatics/btw474

Zhai, J., Jeong, D.-H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., et al. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* 25, 2540–2553. doi: 10.1101/gad.177527.111

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

OPINION
published: 27 May 2020
doi: 10.3389/fgene.2020.00525

Genome-Wide Identification and Characterization of Gene Families in *Arachis*: Methods and Strategies

Yongli Zhang[1], Dongmei Yin[2] and Hui Song[1]*

[1] Grassland Agri-Husbandry Research Center, College of Grassland Science, Qingdao Agricultural University, Qingdao, China, [2] College of Agronomy, Henan Agricultural University, Zhengzhou, China

Keywords: ***Arachis*, gene family, evolution, expression, homology**

To date, at least eight *Arachis* genomes have been completely sequenced, including two *Arachis duranensis*, two *Arachis ipaensis*, one *Arachis monticola*, and three *Arachis hypogaea*. These datasets can provide a powerful starting point to understand the evolution of *Arachis* species. In addition to a comparison of *Arachis* species at the whole-genome level, evolutionary masks can be uncovered based on the analysis of *Arachis* gene families. Although many gene families have been identified and characterized in *Arachis*, different methods and strategies have been used by different researchers. This paper offers advice on the methods and strategies for identification, nomenclature, and quantitative real-time PCR (qRT-PCR) primer-design based on published datasets of *Arachis* gene families. The presented analyses provide a theoretical foundation for the improvement of the identification and characterization of gene families in *Arachis*.

OPEN ACCESS

Edited by:
Ray Ming,
University of Illinois at Urbana-Champaign, United States

Reviewed by:
Cheng Sun,
Chinese Academy of Agricultural Sciences, China
Weilong Hao,
Wayne State University, United States

****Correspondence:***
Hui Song
biosonghui@outlook.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *01 January 2020*
Accepted: *30 April 2020*
Published: *27 May 2020*

Citation:
Zhang Y, Yin D and Song H (2020) Genome-Wide Identification and Characterization of Gene Families in Arachis: Methods and Strategies. Front. Genet. 11:525. doi: 10.3389/fgene.2020.00525

GENOME SEQUENCING AND IDENTIFICATION OF GENE FAMILIES IN *ARACHIS*

The cultivated peanut (*A. hypogaea*, AABB genome) was formed by the crossing of two wild peanuts: *A. duranensis* (AA genome) and *A. ipaensis* (BB genome) (Bertioli et al., 2016, 2019). In 2014, the genome sequences of *A. duranensis* (V14167) and *A. ipaensis* (K30076) were released on PeanutBase (https://peanutbase.org); however, their datasets were not usable at the time because the related paper had not been published then. It was not until 2016 when researchers could begin to use the datasets once the paper was finally published in Nature Genetics (Bertioli et al., 2016). In addition to these two lines, researchers sequenced two other lines: *A. duranensis* (PI 475845) and *A. ipaensis* (ICG_8206) (Chen et al., 2016; Lu et al., 2018). The genome sequences of three cultivated peanut species, namely *A. hypogaea* cv. Tifrunner, *A. hypogaea* cv. Shitouqi, and *A. hypogae*a cv. Fuhuasheng, were sequenced and released in 2018 (Bertioli et al., 2019; Chen et al., 2019; Zhuang et al., 2019). Simultaneously, the genome of a wild tetraploid peanut, *A. monticola*, was completely sequenced (Yin et al., 2018, 2019). These eight available genomic datasets provide raw material for the study of *Arachis* evolution.

Several researchers have focused on genome-wide analyses of the evolution and expression of gene families with canonical domains in *Arachis*. The WRKY transcription factor, a ∼60-residue DNA-binding domain containing a conserved heptapeptide motif WRKYGQK, was first identified after the *A. duranensis* and *A. ipaensis* genomes had been released (Song et al., 2016b). Subsequently, aquaporin (AQP), basic/helix-loop-helix (bHLH), basic leucine zipper (bZIP), EXP (expansin), heat shock transcription factor (HSF), lipoxygenase (LOX), mildew resistance locus (MLO), nucleotide-binding sit–leucine-rich repeat (NBS–LRR), and phosphatidyl ethanolamine-binding protein (PEBP) gene families were identified in the *A. duranensis* (V14167) and *A. ipaensis* (K30076) genomes (Rispail and Rubiales, 2016; Song et al., 2016a, 2017; Gao et al., 2017; Guimaraes et al., 2017; Wang et al., 2017, 2019; Jin et al., 2019; Shivaraj et al., 2019) (**Table S1**).

Growth-regulating factor (GRF) and NBS–LRR gene families were identified in the *A. hypogaea* cv. Tifrunner genome (Song et al., 2019; Zhao et al., 2019) (**Table S1**). However, different methods and strategies were used for the identification of gene families in *Arachis*.

IDENTIFICATION METHOD OF GENE FAMILIES IN *ARACHIS*

At least three methods can be used to identify the members of a gene family. The first method identifies members based on gene annotations. The gene annotation that was generated based on reference genomes was added to the gene name. A gene family was identified using each gene name. This method requires more time when the larger genome is used. In addition, if the gene annotation is wrong, false-positive sequences emerge. The second method identifies members based on local BLAST (PSI-BLAST) or searches tool data from a public database (i.e., PeanutBase). Query sequences always originate from *Arabidopsis thaliana*, *Medicago truncatula*, and *Glycine max*. This method may lose particular gene family members because of species-specific genes. However, this method plays an important role for the identification of gene families with non-canonical domains. The third method identifies members based on a hidden Markov model (HMM) using the HMMER program (Finn et al., 2011). The HMM file was generated by a gene family from various organisms. HMM-based methods can provide an even better representation of gene families and allow the identification of more distant family members.

A total of 12 gene families with canonical domains have been identified in *Arachis* (Rispail and Rubiales, 2016; Song et al., 2016a, 2017, 2019; Gao et al., 2017; Guimaraes et al., 2017; Wang et al., 2017, 2019; Jin et al., 2019; Shivaraj et al., 2019; Zhao et al., 2019). However, researchers used different methods to identify members among these gene families, specifically BLAST-based (four gene families) and HMM-based (eight gene families) methods (**Table S1**). Previous studies have demonstrated that more WRKY gene family members could be identified using the HMM-based method than the BLAST-based method in legumes (Song et al., 2018). To evaluate this result in various *Arachis* gene families, four gene families (AQP, EXP, MLO, and GRF) that were detected using the BLAST-based method in previous studies were re-identified using a HMM-based method. Previous studies identified gene families using different E-value thresholds (**Table S2**). If a smaller E-value was set, a smaller number of gene family members was obtained in the BLAST-based and HMM-based methods. For the PSI-BLAST and HMM programs, the default E-value parameter was 10. To compare the number of identified gene family members that used BLAST-based and HMM-based methods, this study used an E-value of 10 to re-identify the above-mentioned gene family members in *Arachis*. To verify the gene family domain, the obtained sequences were submitted to the Pfam database. The sequence was considered a gene family member if it contained a gene family domain. The obtained results showed that more members were identified using the HMM-based and BLAST-based method with an E-value of 10 than previous studies that used the BLAST-based method with an E-value below 10 among the above-mentioned four gene families (**Figure 1** and **Table S2**). All members from the BLAST-based method were found in the HMM-based methods (**Figure 1** and **Table S3**). In addition to this, compared with the BLAST-based method, the HMM-based method can identify a stable number of gene family members under an E-value of 10 in *Arachis*. Using *A. thaliana*, *Orazy sativa*, and *G. max* AQP and GRF gene family members to query against the *Arachis* genome for identification of a corresponding gene family in BLAST-based method, the same number of gene family members were detected using both the HMM-based and BLAST-based methods in AQP. However, a larger number of gene family members was detected using the HMM-based method than that when the BLAST-based method was used in GRF. In MLO, *A. thaliana* MLO was used as query sequence to identify gene family members in *A. duranensis* and *A. ipaensis*. The results showed that the same number of gene family members was detected using both HMM-based and BLAST-based methods. Nevertheless, more false positive sequences were found in BLAST-based method rather than HMM-based method (**Table S4**). To obtain more gene family members, multiple queries from different plants were considered when the BLAST-based method was used to identify gene families. However, if using the HMM-based method to identify gene families, the query sequence only selected the HMM file. Therefore, the HMM-based method is rapid and accurate. In summary, this study proposes that the best way to identify gene families in *Arachis* is the HMM-based method.

NOMENCLATURE FOR *ARACHIS* GENE FAMILY MEMBERS

The nomenclature for *Arachis* gene family members could be classified into three types (**Table S1**). In the *Arachis* expansin gene family, *A. thaliana* expansin was used as reference (Guimaraes et al., 2017). In brief, the nomenclature for *A. thaliana* expansin was completed based on a chronological order of their discovery and phylogenetic tree (Kende et al., 2004). Synteny was constructed between *Arachis* and *A. thaliana* expansin. In four gene families (bHLH, LOX, and PEBP in *A. duranensis* and *A. ipaensis*; and NBS–LRR in *A. hypogaea* cv. Tifrunner), no nomenclature was allotted for members of gene families. The sequencing ID was used as gene name. In seven gene families (AQP, bZIP, HSF, NBS–LRR, MLO, and WRKY in *A. duranensis* and *A. ipaensis*; and GRF in *A. hypogaea* cv. Tifrunner), the nomenclature for members was defined by their chromosomal order. *Arachis duranensis*, *A. ipaensis*, and *A. hypogaea* cv. Tifrunner were referred to as Ad, Ai, and Ah, respectively. Following this procedure, the gene family name was listed and the number was then assigned based on the gene location in chromosomal order (e.g., AdWRKY1 and AdWRKY2). However, if a new member was found after the nomenclature had been assigned to a given gene family, the gene order of the new member should come after the last number of the legacy version.

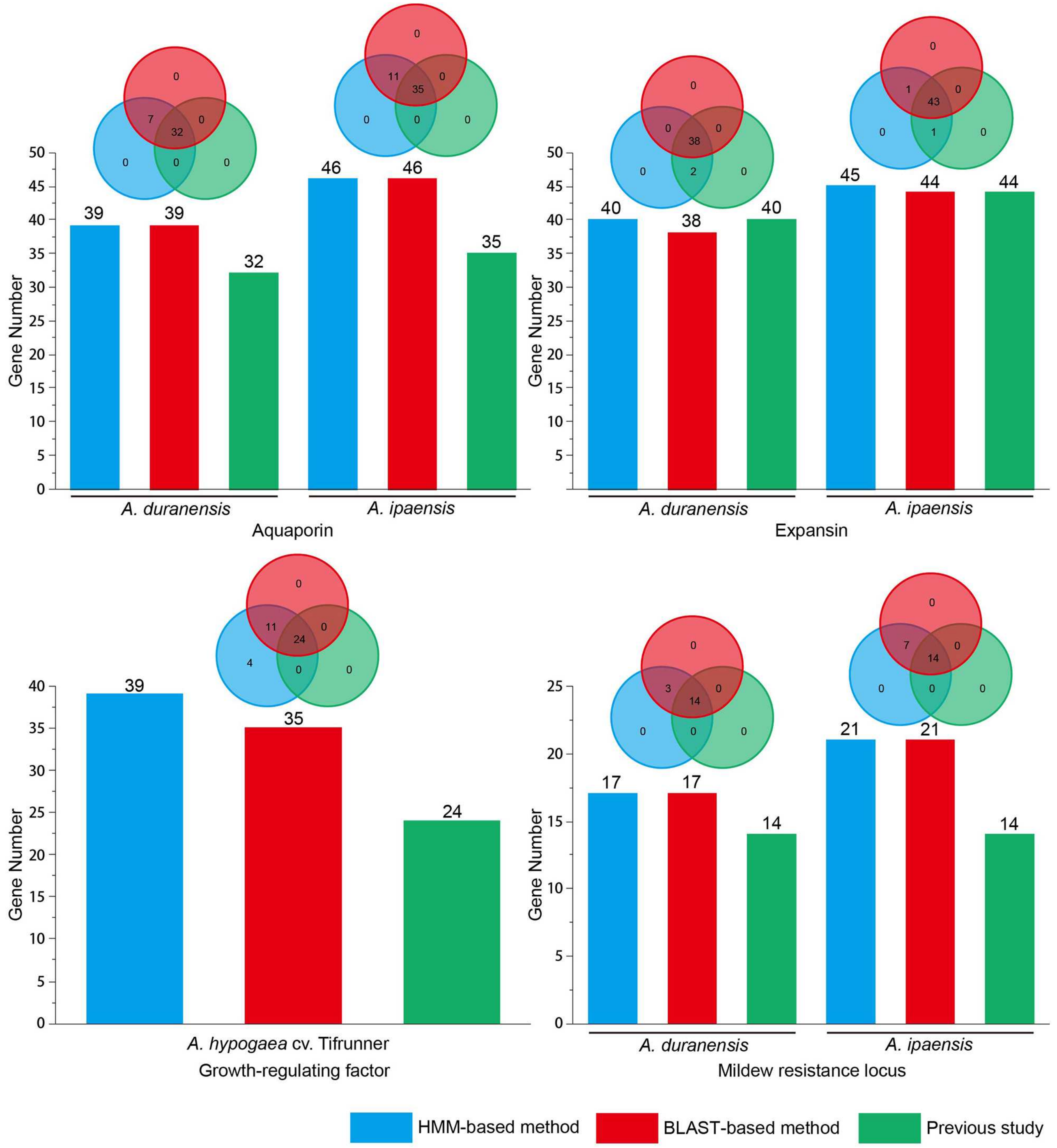


FIGURE 1 | Identification of gene families in *Arachis* using HMM-based and BLAST-based method. Previous aquaporin data from Shivaraj et al. (2019); Previous expansin data from Guimaraes et al. (2017); Previous growth-regulating factor data from Zhao et al. (2019); Previous mildew resistance locus data from Rispail and Rubiales (2016).

IDENTIFICATION OF DUPLICATED GENES IN *ARACHIS* GENE FAMILIES

Gene duplication is one of the driving forces of evolution and is a potential strategy for the adaptation to environmental change (Panchy et al., 2016; Van de Peer et al., 2017). To date, nine gene families were used to conduct homolog (paralog and ortholog) relationship analysis (**Table S1**). However, different methods were used to identify homology in *Arachis*, including phylogenetic tree, BLAST-based methods, and synteny relationship methods (Rispail and Rubiales, 2016; Song et al., 2016a,b, 2017, 2019; Guimaraes et al., 2017; Wang et al., 2017, 2019; Jin et al., 2019). Although these methods have been used to identify homologs in many studies, detailed parameters need to be listed. For example, which model was used and which bootstrap was credible for clades in the phylogenetic tree? Which threshold value was set for the synteny analyses? This paper recommends that researchers should consider using the BLAST-based homolog identification method in *Arachis* because this method has been verified for the identification of homologs in the cultivated peanut (Clevenger et al., 2016; Bertioli et al., 2019; Chen et al., 2019; Zhuang et al., 2019). The following evaluation criteria were used as thresholds to determine homology: (1) alignment coverage exceeding 80% of the two sequences; (2) identity $>$ 80%; and (3) E-value $\leq$ 1E−10.

Gene completeness is a crucial factor that affects evolutionary analysis. Confusing results can be obtained when partial sequences are used in gene structure analysis because of the potential loss of introns and exons. In addition, selection pressure cannot be identified when partial homolog sequences are used. Therefore, it is suggested that full-length sequences of *Arachis* gene family members should be used for the evolutionary analyses. In addition to this, it is also worth noting that pseudogenes were identified during analysis of gene families. Although pseudogenes may play a crucial role in plant development and response to stress, most pseudogenes cannot code for proteins or loss of the original function. Therefore, pseudogenes were excluded when the selective pressures were estimated. In *A. duranensis* and *A. ipaensis*, CDSs with premature codons were reported in MLO, NBS–LRR, and WRKY gene families, which have been considered pseudogenes (Rispail and Rubiales, 2016; Song et al., 2016b, 2017).

QRT-PCR PRIMER DESIGN FOR *ARACHIS* GENE FAMILIES

The cultivated peanut is allotetraploid and contains many homologs. In addition, the members of gene families contain conserved sequences. Therefore, qRT-PCR primers are difficult to design because of non-specific amplification. Before the cultivated peanut genome was released, qRT-PCR primers were designed using the sum of *A. duranensis* and *A. ipaensis* sequences as cultivated peanut genome (Song et al., 2016a, 2017). Researchers focused on a problem to avoid the amplification of homologous sequences when designing the qRT-PCR primers in *Arachis* NBS–LRR and LOX gene families (Song et al., 2016a, 2017). Until now, the cultivated peanut genome can be used to study the expression of gene families. Future study has to carefully design the qRT-PCR primers to avoid non-specific amplification. The qRT-PCR primers are designed using the CDS with untranslated region (UTR) sequence because the UTR contained non-conserved sequences. Non-conserved regions are identified using multiple sequence alignment before designing the qRT-PCR primers. The Beacon Designer program was used for designing qRT-PCR primers. Beacon Designer can upload the genome sequence as a database. When a pair of qRT-PCR primers is designed, the program searches the database and lists the amplified fragment. This function can help researchers to remove false-positive primers.

CONCLUSIONS

With the released *Arachis* genome sequence, more gene families can be identified and characterized. This study offers advice on gene family identification and characterization in *Arachis*. The HMM-based method can be used to identify members of a given gene family. Full-length sequences were used for evolutionary analysis. Homologs can be identified by a BLAST-based method. Non-specific amplification can be avoided in qRT-PCR.

AUTHOR CONTRIBUTIONS

HS and YZ conceived the study. HS wrote the paper. HS and DY approved the final version.

FUNDING

This work was supported by the First Class Grassland Science Discipline Program of Shandong Province, China, and the Natural Science Foundation of Shandong Province, China (ZR2019QC017).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00525/full#supplementary-material

Table S1 | Identification of 12 gene families in *Arachis*.

Table S2 | Comparison of four *Arachis* gene families using HMM-based and BLAST-based methods. ^{a}The E-value was set to 10 in both the HMM-based and BLAST-based methods.

Table S3 | Gene names in the four *Arachis* gene families. ^{a}The E-value was set to 10 in both the HMM-based and BLAST-based methods.

Table S4 | False positive rates in HMM-based and BLAST-based methods.

REFERENCES

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., et al. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* 51, 877–884. doi: 10.1038/s41588-019-0405-z

Chen, X., Li, H., Pandey, M. K., Yang, Q., Wang, X., Garg, V., et al. (2016). Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6785–6790. doi: 10.1073/pnas.1600899113

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., et al. (2019). Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant.* 12, 920–934. doi: 10.1016/j.molp.2019.03.005

Clevenger, J., Chu, Y., Scheffler, B., and Ozias-Akins, P. (2016). A developmental transcriptome map for allotetraploid *Arachis hypogaea*. *Front. Plant. Sci.* 7:1446. doi: 10.3389/fpls.2016.01446

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Gao, C., Sun, J., Wang, C., Dong, Y., Xiao, S., Wang, X., and Jiao, Z. (2017). Genome-wide analysis of basic/helix-loop-helix gene family in peanut and assessment of its roles in pod development. *PLoS ONE* 12:e0181843. doi: 10.1371/journal.pone.0181843

Guimaraes, L. A., Mota, A. P. Z., Araujo, A. C. G., de Alencar Figueiredo, L. F., Pereira, B. M., de Passos Saraiva, M. A., et al. (2017). Genome-wide analysis of expansin superfamily in wild *Arachis* discloses a stress-responsive expansin-like B gene. *Plant Mol. Biol.* 94, 79–96. doi: 10.1007/s11103-017-0594-8

Jin, H., Tang, X., Xing, M., Zhu, H., Sui, J., Cai, C., and Li, S. (2019). Molecular and transcriptional characterization of phosphatidyl ethanolamine-binding proteins in wild peanuts *Arachis duranensis* and *Arachis ipaensis*. *BMC Plant Biol.* 19:484. doi: 10.1186/s12870-019-2113-3

Kende, H., Bradford, K. J., Brummell, D. A., Cho, H. T., Cosgrove, D. J., Fleming, A. J., et al. (2004). Nomenclature for members of the expansin superfamily of genes and proteins. *Plant Mol. Biol.* 55, 311–314. doi: 10.1007/s11103-004-0158-6

Lu, Q., Li, H., Hong, Y., Zhang, G., Wen, S., Li, X., et al. (2018). Genome sequencing and analysis of the peanut B-genome progenitor (*Arachis ipaensis*). *Front. Plant. Sci.* 9:604. doi: 10.3389/fpls.2018.00604

Panchy, N., Lehti-Shiu, M., and Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171, 2294–2316. doi: 10.1104/pp.16.00523

Rispail, N., and Rubiales, D. (2016). Genome-wide identification and comparison of legume MLO gene family. *Sci. Rep.* 6:32673. doi: 10.1038/srep32673

Shivaraj, S. M., Deshmukh, R., Sonah, H., and Bélanger, R. R. (2019). Identification and characterization of aquaporin genes in *Arachis duranensis* and *Arachis ipaensis* genomes, the diploid progenitors of peanut. *BMC Genomics* 20:222. doi: 10.1186/s12864-019-5606-4

Song, H., Guo, Z., Hu, X., Qian, L., Miao, F., Zhang, X., and Chen, J. (2019). Evolutionary balance between LRR domain loss and young NBS-LRR genes production governs disease resistance in *Arachis hypogaea* cv. *Tifrunner. BMC Genom.* 20:844. doi: 10.1186/s12864-019-6212-1

Song, H., Sun, W., Yang, G., and Sun, J. (2018). WRKY transcription factors in legumes. *BMC Plant Biol.* 18:243. doi: 10.1186/s12870-018-1467-2

Song, H., Wang, P., Li, C., Han, S., Lopez-Baltazar, J., Zhang, X., et al. (2016a). Identification of lipoxygenase (LOX) genes from legumes and their responses in wild type and cultivated peanut upon Aspergillus flavus infection. *Sci. Rep.* 6:35245. doi: 10.1038/srep35245

Song, H., Wang, P., Li, C., Han, S., Zhao, C., Xia, H., et al. (2017). Comparative analysis of NBS-LRR genes and their response to *Aspergillus flavus* in *Arachis*. *PLoS ONE* 12:e0171181. doi: 10.1371/journal.pone.0171181

Song, H., Wang, P., Lin, J. Y., Zhao, C., Bi, Y., and Wang, X. (2016b). Genome-wide identification and characterization of *WRKY* gene family in peanut. *Front Plant Sci.* 7:534. doi: 10.3389/fpls.2016.00534

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

Wang, P., Song, H., Li, C., Li, P., Li, A., Guan, H., et al. (2017). Genome-wide dissection of the heat shock transcription factor family genes in *Arachis*. *Front Plant Sci.* 8:106. doi: 10.3389/fpls.2017.00106

Wang, Z., Yan, L., Wan, L., Huai, D., Kang, Y., Shi, L., et al. (2019). Genome-wide systematic characterization of bZIP transcription factors and their expression profiles during seed development and in response to salt stress in peanut. *BMC Genom.* 20:51. doi: 10.1186/s12864-019-5434-6

Yin, D., Ji, C., Ma, X., Li, H., Zhang, W., Li, S., et al. (2018). Genome of an allotetraploid wild peanut *Arachis monticola*: a de novo assembly. *GigaScience* 7:giy066. doi: 10.1093/gigascience/giy066

Yin, D., Ji, C., Song, Q., Zhang, W., Zhang, X., Zhao, K., et al. (2019). Comparison of *Arachis monticola* with diploid and cultivated tetraploid genomes reveals asymmetric subgenome evolution and improvement of peanut. *Adv. Sci.* 28:1901672. doi: 10.1002/advs.201901672

Zhao, K., Li, K., Ning, L., He, J., Ma, X., Li, Z., et al. (2019). Genome-wide analysis of the growth-regulating factor family in peanut (*Arachis hypogaea* L.). *Int. J. Mol. Sci.* 20:4120. doi: 10.3390/ijms20174120

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* 51, 865–876. doi: 10.1038/s41588-019-0402-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 20 July 2020
doi: 10.3389/fgene.2020.00747

Identification of Ear Morphology Genes in Maize (*Zea mays* L.) Using Selective Sweeps and Association Mapping

Ting Li[1,2†], Jianzhou Qu[1,2†], Xiaokang Tian[1,2], Yonghui Lao[1,2], Ningning Wei[1,2], Yahui Wang[1,2], Yinchuan Hao[1,2], Xinghua Zhang[1,2], Jiquan Xue[1,2] and Shutu Xu[1,2*]*

[1] The Key Laboratory of Biology and Genetics Improvement of Maize in Arid Areas of the Northwest Region, Ministry of Agriculture, College of Agronomy, Northwest A&F University, Xianyang, China, [2] The Maize Engineering Technology Research Centre of Shaanxi Province, Yangling, China

Edited by:
Yin Li,
Rutgers, The State University
of New Jersey, United States

Reviewed by:
Chenxu Liu,
China Agricultural University, China
Rohit Kumar,
Clemson University, United States

***Correspondence:**
Jiquan Xue
xjq2934@163.com
Shutu Xu
shutuxu@nwafu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *06 March 2020*
Accepted: *23 June 2020*
Published: *20 July 2020*

Citation:
Li T, Qu J, Tian X, Lao Y, Wei N, Wang Y, Hao Y, Zhang X, Xue J and Xu S (2020) Identification of Ear Morphology Genes in Maize (Zea mays L.) Using Selective Sweeps and Association Mapping. Front. Genet. 11:747. doi: 10.3389/fgene.2020.00747

The performance of maize hybrids largely depend on two parental inbred lines. Improving inbred lines using artificial selection is a key task in breeding programs. However, it is important to elucidate the effects of this selection on inbred lines. Altogether, 208 inbred lines from two maize heterosis groups, named Shaan A and Shaan B, were sequenced by the genotype-by-sequencing to detect genomic changes under selection pressures. In addition, we completed genome-wide association analysis in 121 inbred lines to identify candidate genes for ear morphology related traits. In a genome-wide selection scan, the inbred lines from Shaan A and Shaan B groups showed obvious population divergences and different selective signals distributed in 337 regions harboring 772 genes. Meanwhile, functional enrichment analysis showed those selected genes are mainly involved in regulating cell development. Interestingly, some ear morphology related traits showed significant differentiation between the inbred lines from the two heterosis groups. The genome-wide association analysis of ear morphology related traits showed that four associated genes were co-localized in the selected regions with high linkage disequilibrium. Our spatiotemporal pattern and gene interaction network results for the four genes further contribute to our understanding of the mechanisms behind ear and fruit length development. This study provides a novel insight into digging a candidate gene for complex traits using breeding materials. Our findings in relation to ear morphology will help accelerate future maize improvement.

Keywords: population characteristics, selective sweeps, genome-wide association study, co-expression, ear morphology

INTRODUCTION

Maize (*Zea mays* L. ssp. *mays*), one of the most widely grown crops in the world, plays an essential role in global food security. It has been suggested that maize was domesticated from teosinte (*Zea mays* L. ssp. *Parviglumis*) about 10,000 years ago in the Balsas River Basin of southwestern Mexico (Schnable et al., 2009; Xiao et al., 2017). Originally, the teosinte was a wild plant with approximately 5 to 12 kernels in the ear and long tassel branches (Doebley et al., 1995; Wang et al., 1999). The number of kernels in modern maize varieties is far larger than that of teosinte and the long branches have disappeared. There are sharp distinctions between modern maize and teosinte in terms of plant morphology structure.

The morphological architecture of maize underwent a striking transformation mainly owing to selection by early agriculturalists and the local environment. This domestication stage continued for a long time. The original landraces appeared during this long-term domestication process (Yamasaki et al., 2005). In recent years, to increase production and quality levels, modern breeders have made great improvements to maize materials. Recent short-term artificial improvements of maize have resulted in the improvement of combining ability, including specific combining abilities and others (Souza et al., 2010; Viana et al., 2013). Besides phenotypic change, selection during the domestication and improvement processes also brought about reduced genetic diversity in selected genes (Yamasaki et al., 2005). In maize, diversity analyses showed a sequence diversity decrease in the promoter region of *Teosinte branched1* (*tb1*) which affects long branches (Wang et al., 1999). Analysis of the genomic history of North American maize found that genetic separation and linkage disequilibrium increased with maize improvement (Van Heerwaarden et al., 2012). Selection transformed the genome and the shape of maize during domestication and improvement (Shi and Lai, 2015). Such changes in the genome and phenotype during the selection process provide an opportunity to study the influence of selected genes on agronomic traits.

A number of studies have explored the genetic variants of selected traits that arose during the selection process in various species. In soybeans, more than 800 differentiated regions, selected during the domestication process, were observed in the genome (Han et al., 2016). In chrysanthemum, about 550 genomic regions underwent selection amongst the different chrysanthemum types (Chong et al., 2017). In rice, a total of 200 regions with selective signatures were detected in two major *indica* subpopulations. A large number of genes with known functions in these selected regions are associated with crucial agronomic traits, particularly grain yield (Xie et al., 2015). In maize, there are 484 selected features arising from domestication and 695 from improvement (Hufford et al., 2012). In fact, the selected traits in different breeding populations are diverse during maize improvement. Thus, identifying the genetic variants underlying agronomic traits in different populations, which were developed during crop improvement, will help us understand more about selection effects and, in turn, lead to further crop improvements.

With the advent of new technologies and the exploitation of diverse analysis approaches, it is now possible to identify important genetic variants related to selected traits. Selective sweeps have been used to dissect selected regions in the above-mentioned studies. In general, the selection of a target trait is always accompanied by other agronomic traits due to the genetic hitch-hiking effect (Xu, 2018). The genome-wide association study (GWAS) has also been widely used to identify loci linked to target traits (Li et al., 2013; Chen et al., 2018). GWAS has a higher resolution to obtain casual genes, but some will be false positive. Hence, combining selective sweeps and GWAS is an efficient way to identify selected candidates in relation to specific traits. Furthermore, gene co-expression networks can be set up using gene expression data, which help associate genes of unknown function with biological processes and prioritize candidate select/target genes or discern transcriptional regulatory programs. This process has been carried out on data from multiple plants and has illuminated many key events during plant development (Sarkar et al., 2014; Huang et al., 2017; Wisecaver et al., 2017; Yu et al., 2017). Each approach has its unique advantages and disadvantages. Therefore, combining multiple analysis methods is likely to be an effective way of identifying the genetic mechanisms of selected traits.

Traits relating to ear architecture such as ear length (EL), fruit length (FL), setting rate (SR), and barren tip length (BTL) are essential for the improvement of grain yield during the breeding process. Two heterotic groups (Shaan A group and Shaan B group), with high heterosis between them, were established during 10-year breeding programs. Usually, inbreds from the Shaan A group act as the female population and those from the Shaan B group act as male. These sex differences have been utilized during selection to increase combining ability. In this study, we found that some traits arising from lines developed during long-term breeding selection from Shaan A and Shaan B groups showed significant differences, particularly in relation to EL, FL, BTL, and SR. This provides a good resource for the study of genetic variants in EL, FL, BTL, and SR. Here, we identified massive selective regions in the inbred lines from Shaan A and Shaan B groups, and we dissected the genetic mechanisms of EL, FL, BTL, and SR using this breeding population. In addition, we developed global gene co-expression networks and spatiotemporal-specific processes of genes in selected regions and genes governing significant agronomic traits. Our research revealed some candidate genes associated with ear development, which provides a valuable data resource for plant genetics research and breeding.

MATERIALS AND METHODS

Plant Materials and SNP Genotyping

A total of 208 maize inbred lines (AM208) were collected including 54 inbred lines from the Shaan A group (A54) and 154 inbred lines from the Shaan B group (B154) (**Supplementary Table S1**). The Shaan A and Shaan B groups are derived from one basic population which was constructed using several excellent varieties. Then, the elite inbred Ye478 (Reid group) and HuangZaoSi (Tang SPT group) from China were used to pull the basic population into Shaan A group and Shaan B group, which were then improved over a period of approximately 10 years. Within the two heterosis groups, Shaan A and Shaan B play the role as female and male, respectively. The materials in the two groups were selected based on the number of harvest ears, grain weight per ear and seed rate (seed weight/ear weight) by planting in high density, low nitrogen and low irrigation conditions. New inbreds can be filtered by conducting a combining ability test which crosses more than two elite inbreds (including Zheng58 and Chang7-2). Two hundred and eight inbred lines were sampled at the 3-leaf stage and the DNA was extracted using a modified CTAB method (Murray and Thompson, 1980). The genotypes were determined using tGBS technology (Dara2bio; LLC, Ames, IA, United States) (Li et al., 2018). Overall, 48,432

SNPs were retained. Later, the 48,432 SNPs were inputted into Beagle version 4.1 analysis software (Browning and Browning, 2007, 2016) and filtered by a minor allele frequency (MAF) cutoff point of 5% using PLINK version 1.90 software (Purcell et al., 2007). A set of 32,306 high-quality SNPs (MAF $\geq$ 0.05) was retained for further analysis. In addition, 121 lines were selected from the AM208 population for construction of an association population (AM121), of which 26 lines (A26) belong to A54 and 95 (B95) belong to B154. Genotype data from the association population were filtered from the AM208 population genotype file resulting in 32,306 SNPs which were further screened using a MAF ($\geq$0.05). Ultimately, 32,051 high-quality SNPs were applied to the GWAS.

Population Structure and LD Analyses

For the AM208, an UPGMA tree with 1000 bootstraps was constructed using MEGA 7.0 software and a set of 32,306 high-quality SNPs (Kumar et al., 2016). Principal component analysis (PCA) was performed using the GCTA tool on the high-quality SNPs (Yang et al., 2011). The output of the GCTA tool was inputted into R software so as to graphically display the PCA results. Additionally, A54 and B154 were separated to compute the linkage disequilibrium (LD) decay distance of each group. To avoid the effect of sample size, 54 inbreds (B54 re-sample) were selected from B154 100 times at random. The LD decay distance was calculated using PopLDdecay software (Zhang et al., 2019). For AM121, a kinship matrix was created and a PCA was performed using TASSEL software version 5.0 (Bradbury et al., 2007) and 32,051 SNPs (MAF $\geq$ 0.05).

Screening for Selective Regions and Genome-Wide Associations (GWAS)

Selective regions typically include two features, high differentiation and low diversity (Doebley et al., 2006). Therefore, the nucleotide diversity (π) ratios (A54/B154), the genetics differentiation (*Fst*) and the Tajima's *D* statistic were calculated using a sliding-windows approach (100 kb non-overlapping sliding window) for A54 and B154 (Schmutz et al., 2014). We identified the low diversiry regions in A54 by the bottom 10% πA54/πB154 value and in B154 by the top 10% πA54/πB154 value. When a window is located on both the top 10% of the pool's empirical distribution for *Fst* and the low diversity regions in A54 or B154, the window is considered to be a selective region in either the Shaan A or the Shaan B group.

Phenotypic data and 32,051 SNPs from AM121 were used to uncover the genetic architecture of the target traits using GWAS and a linear mixed model (P + K) in TASSEL v.5.0 software. The threshold was set to 1×10^{-3} based on analysis results. When the associated regions of the significant SNPs relating to the target trait overlapped with the selected regions, these associated regions were considered to be candidate regions and subsequent analysis was performed. In these regions, all genes were identified using the MaizeGDB database[1]. To compare selected genes with those from published data, all v4 selected gene IDs were converted to v3 gene IDs.

Field Experiment Design and Phenotypic Data Collection

For the association population, AM121 individuals were planted in two different fields, which were distinguished using the codes E1 and E2, in Yangling in Shaanxi Province, China in 2017. The experiment consisted of two replicates, each with two rows. Each row was 5 m long and 0.6 m wide. The planting density was 67,500 plants/ha. When the maize was mature, according to single ear weight, ten ears were selected in each plot in the first replicate to measure ear weight (EW, g), and five ears were selected to measure ear row number (ERN), ear diameter (ED, cm), kernel number per row (KNR), ear length (EL, cm), fruit length (FL, cm), barren tip length (BTL, cm) and the compute setting rate (SR, the ratio of FL to EL). The mean value per material for each parameter was calculated. Descriptive statistics, ANOVAs, and Pearson correlation analyses were conducted using SPSS v.22 software (IBM crop. Armonk, NY, United States). Broad-sense heritability (H^2) was calculated as follows:

$$H = \frac{\sigma_g}{\sigma_g + \frac{\sigma_e}{k}},$$

where ${\sigma_g}^2$ is the genetic variance, ${\sigma_ç}^2$ represents error variance, and k is the number of environments (Hallauer and Miranda, 1981).

Gene Co-expression Network Analysis

A total of 78 maize B73 RNA-seq datasets from multiple tissues (seed, endosperm, embryo, leaf, ear, tassel, silk, cob, root, shoot, pollen, anther, and SAM) and developmental stages, reported and available in a public database[2], were used to construct gene co-expression networks (Chen et al., 2014; Li et al., 2014). Gene co-expression network analysis was performed using the R package WGCNA version 1.63 (Langfelder and Horvath, 2008). In addition, β was optimized to six to achieve a scale-free topology. Hierarchical clustering was used to identify gene modules with a dynamic tree-cutting algorithm based on the dissimilarity of gene connectivity. A NetworkAnalyzer plugin available in Cytoscape was used to calculate relevant network parameters, such as the degree of connection (Assenov et al., 2008). Modules were visualized using Cytoscape version 3.6.1.

Functional Annotation of Genes

To obtain more complete annotation information, all protein sequences of maize were mapped to Swiss-Prot/UniProt, Pfam, InterPro, Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases by eggnog-mapper. For a given gene set, the R package clusterprofiler was used to visualize GO terms (Yu et al., 2012). A GO term was considered significantly enriched if the adjusted *p*-value was lower than 0.05. The pathway maps were obtained from the KEGG database (Kanehisa et al., 2017). The pathway enrichment analysis was performed using KOBAS version 2.0, and a Benjamini and Hochberg adjusted *p*-value of 0.05 was used as the cut-off criterion (Xie et al., 2011).

[1] https://www.maizegdb.org/gbrowse/maize_v4

[2] https://maizegdb.org/

Expression and Statistical Analysis

To avoid an infinite value, gene values expressed as zero were replaced with a value of 0.01. All of the samples were normalized using $\log_2$ (expression values + 0.01). Hierarchical clustering analysis was performed using the R package pheatmap[3] and Pearson's correlation co-efficient as the distance measure. Additionally, a Venn diagram was drawn using the VennDiagram package in R (Chen and Boutros, 2011).

RESULTS

Genetic Divergence Between Inbred Lines From Shaan A and Shaan B Groups

In order to enlarge the germplasm resource of maize and screen elite inbred lines, the two heterosis groups (Shaan A group and Shaan B group) were built and improved. A series of inbred lines from both groups were selected for breeding and genetic study during improvement. To reveal genetic divergence between inbred lines from the Shaan A and Shaan B groups at the genomic level, principal component analysis (PCA) and an UPGMA tree were conducted using 32,306 high-quality SNPs with a MAF greater than 0.05. In the PCA, most of the inbred lines from the Shaan A group were separated from the Shaan B group lines using the first two or three eigenvectors (**Figures 1A,B**). Inbreds from Shaan B group were dispersed. In the UPGMA tree, except for several inbreds, inbred lines from the Shaan A group also clustered together, as did inbred lines from the Shaan B group (**Figure 1C**). This illustrated that Shaan A group and Shaan B group have occurred divergence in the genome, but some inbreds from the Shaan A and Shaan B groups are more closely related. Therefore, continuous genetic improvement in the future is needed to achieve further population divergence. Furthermore, in terms of the LD decay distance (r^2 reached 0.2), AM208 was approximately 10 kb, and A54 had a longer LD decay distance even when we compared it to results of the same number of inbred lines when based on random sampling from B154. This indicates that B154 has a higher genetic diversity than A54 (**Figure 1D**).

To identify the selective regions, the *Fst* and π ratio values between A54 and B154 were calculated using a 100kb step length. A54 had a lower nucleotide diversity (median, πA54/πB154 = 0.854) than B154. In addition, 91 selected regions (9.1 Mb) with both low diversity and high differentiation were discovered in A54, which included 199 protein-coding genes. However, in B154, a total of 246 (24.6 Mb) selected regions were identified, which contained 573 protein-coding genes (**Figures 2A,B** and **Supplementary Table S2**). The number of genomic regions with selective sweep signals in B154 was approximately 2.7 times that of A54. Moreover, selective regions in A54 focus on chromosome 3 and 7 (48, 52.75%), but are generally distributed on 10 chromosomes in B154. In addition, of these selective regions, 15 in A54 and 38 in B154 have no genes.

Overview of the Functional Diversity of Genes and the Co-expression Network in Selected Regions

To investigate the functional genes located in selected regions, GO analysis was performed. There were distinctive differences in the greatly enriched GO terms ($p < 0.05$) of selected genes between A54 and B154. Significantly, genes in the selected regions for B154 showed greater diversity in biological functions than those in A54, such as in hormone synthesis, cell development, and partial secondary metabolites (**Supplementary Table S3**). In addition, the selective regions located on specific chromosomes and the biological functional diversity of genes in different groups indicate that there is a chasm-like split in some inbred line phenotypes from the Shaan A and Shaan B groups.

To gain a deeper understanding of the effect of genes in the selected regions, a weighted gene co-expression network analysis (WGCNA) was performed using gene expression data from multiple tissue types and developmental stages, which is available from a public database (see text footnote 2). A total of 23 WGCNA modules were identified in the analysis (**Figure 3A**). Of these, the green module was the most distinct (**Figures 3B,C**). Notably, genes in the selected regions of A54 and B154 were mainly distributed in 18 modules except for genes not included in the WGCNA (**Figure 3D**). To make it easier to follow the relationships among genes in selected regions and other unselected genes, 199 genes from A54 and 573 genes from B154, in selected regions, were screened as potential co-expression genes based on a weight value higher than 0.6. A co-expression network was then constructed (**Figure 4**). In total, 1086 genes were found to interact with selected genes from A54 and B154. In addition, 452 genes interacted with selected genes from A54 and 1315 genes interacted with selected genes from B154 (**Supplementary Table S4**). GO enrichment analysis indicated that the common interactive genes from A54 and B154 were mainly enriched during tissue development, regulation from vegetative to reproductive states, and pollen germination. Furthermore, remarkable functional differences were noted between specific genes in the selected regions of A54 and B154 (**Supplementary Figure S1** and **Supplementary Table S5**). These results not only show new non-described gene associations and allow the placement in a functional context of some unknown non-assigned genes based on their interactions with known gene families, but also show that improvement for maize may has further accelerated the polarization of A54 and B154 in changing certain maize characteristics (such as quality and resistance) to satisfy breeding requirements.

Genomic Association Analysis for Ear Morphology Components

Following functional enrichment analysis of the interactive genes, we found that these genes were mainly enriched during tissue development processes. During inbred selection, breeders give full attention to morphology and genetic diversity changes due to their importance for hybridization. Herein, we chose eight ear related traits containing ear length (EL), fruit length (FL), ear weight (EW), ear row number (ERN), ear diameter (ED), kernel

[3] https://cran.r-project.org/web/packages/pheatmap/index.html

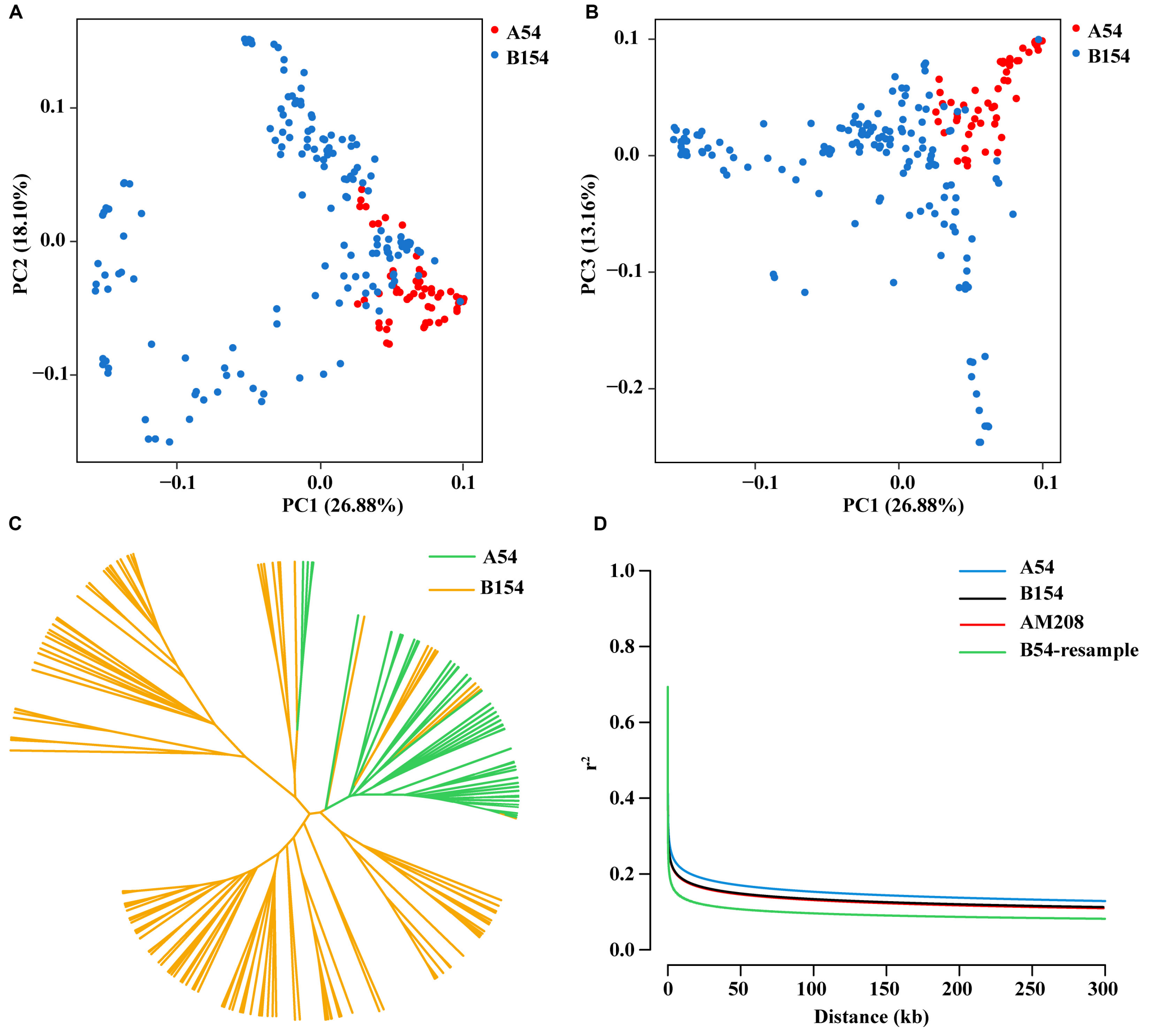


FIGURE 1 | Population characteristic of AM208. **(A)** Scatter plot for PC1 and PC2. **(B)** Scatter plot for PC1 and PC3. **(C)** Phylogenetic tree of AM208 was constructed with 32,306 SNPs. **(D)** Linkage disequilibrium (LD) decay of AM208, A54, and B154.

number per row (KNR), barren tip length (BTL) and setting rate (SR) in order to observe phenotypic differences between 26 inbred lines (A26) from the Shaan A group and 95 inbred lines (B95) from the Shaan B group. Except for ERN and ED, there were significant differences in EL, FL, EW, SR, BTL, and KNR between A26 and B95 in the two environments (**Figure 5**). In addition, EL was significant positively correlated with FL (r = 0.95). However, SR and BTL were significant negatively correlated (**Supplementary Figure S2**). These results indicate that ear related traits were significantly altered in the Shaan A and Shaan B groups during population improvement process.

In the AM121 population, the broad-sense heritability (H^2) of the eight ear traits were higher than 60%, ranging from 65.33% for BTL to 79.09% for EW (**Table 1**). Considering the observed phenotypic differences between A26 and B95 and the correlation coefficient between traits, EL, FL, BTL and SR were chosen to help conduct the GWAS with 32,051 high-quality SNPs, analyzed using a linear mixed model (P + K). As shown in **Supplementary Figure S3**, the probability of a false positive result has been reduced in this study. In total, 84, 49, 47, and 50 significant SNPs were identified for EL, FL, SR, and BTL, respectively (**Figure 6** and **Supplementary Table S6**). Among these SNPs, 5, 4, 3, and 0 SNPs in relation to EL, FL, BTL and SR were co-localized in two locations,respectively. Furthermore, we noticed that the majority of significant SNPs were co-localized between EL and FL, as were BTL and SR. Especially, three co-localized SNPs associated with EL and FL located at chromosome 1 were observed with the same p-value due to the high LD (r^2 = 1.0) (**Figures 2C–F**). In relation to these three significant SNPs, two different haplotypes (GGG, AAA) were identified. Haplotype 2 (AAA) had a longer EL and FL. Approximately 95.79% (91) of B95 was associated with Haplotype 1 (GGG) and 38.46% (10) of A26 was associated with

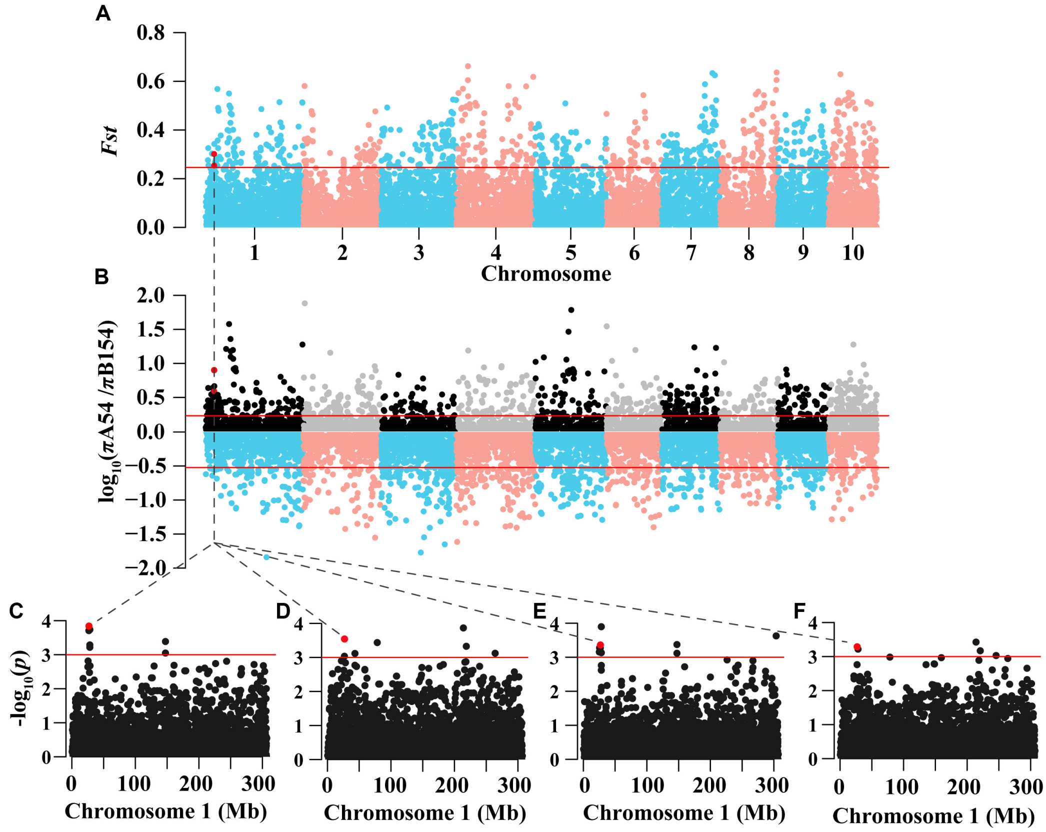


FIGURE 2 | Genomic regions through long-term artificial selection in AM208 and Manhattan plot of EL and FL in Chromosome 1. **(A)** Distribution of genetics differentiation (*Fst*) values. **(B)** Distribution of nucleotide diversity (π) ratio (πA54/πB154). These values are calculated in 100kb sliding windows. Red lines represent the 90% tails of the empirical distribution of each stastic. **(C)** Region plot of EL in E1. **(D)** Region plot of EL in E2. **(E)** Region plot of FL in E1. **(F)** Region plot of FL in E2.

Haplotype 2 (AAA) (**Supplementary Table S7**). In addition, none of co-localized significant SNPs were identified in relation to BTL and SR in both environments. The reason for this may be that BTL and SR have a lower heritability than EL and FL. According to the LD decay distance reported in previous research, the upstream and downstream 150 kb of the significant SNPs are regarded as associated regions (Li et al., 2018). Associated regions including the same genes were considered as one associated region for each trait. Finally, 404, 182, 207, 154 protein-coding genes (reference maize genome v4) distributed in these associated regions were found to be associated with EL, FL, BTL, and SR, respectively (**Supplementary Table S6**).

Gene Co-location: Combining Selective Sweeps and Association Analysis

By integrating selective sweeps and GWAS, we identified 5 regions related to EL or FL and another 5 associated regions related to BTL or SR overlapping with the selected regions (**Table 2**). There was a difference between the size of the selected window and the associated region. Therefore, the 10 associated regions were designated as candidate regions so as to avoid missing relevant genes. Interestingly, among these regions, the candidate region (26,738,827–27,038,887 bp), located on chromosome 1, was from the three co-localized SNPs related to EL and FL (above-mentioned) (**Figures 2C–F**). According to information from the reference genome sequence (www.maizegdb.org_v4), four protein-coding genes were found to be distributed in this 300 kb candidate region, which contains *Zm00001d028216*, *Zm00001d028217*, *Zm00001d028218*, and *Zm00001d028219*.

Of the four candidate genes, *Zm00001d028216* was annotated as *indeterminate floral apex1* and synonyms with a C2C2-YABBY transcription factor, which has been found to be involved in regulating the determinacy of the floral meristem, spikelet pair meristems and spikelet meristems (Laudenciachingcuanco and Hake, 2002). The second gene, *Zm00001d028217*, was annotated as the developmental protein SEPALLATA 2 and descripted as a MADS transcription factor-*MADS14*. It has a wide range of functions, such as regulating flowering time, vegetative development and fruit ripening, especially in the meristem, and is related to floral organ identity (Cacharr et al., 1999; Heuer et al., 2001; Setter et al., 2001; Danilevskaya et al., 2008; Thompson and Hake, 2009; Zhang et al., 2012). The other two genes may have essential functions in defining the boundary of secondary walls and are described as follows:

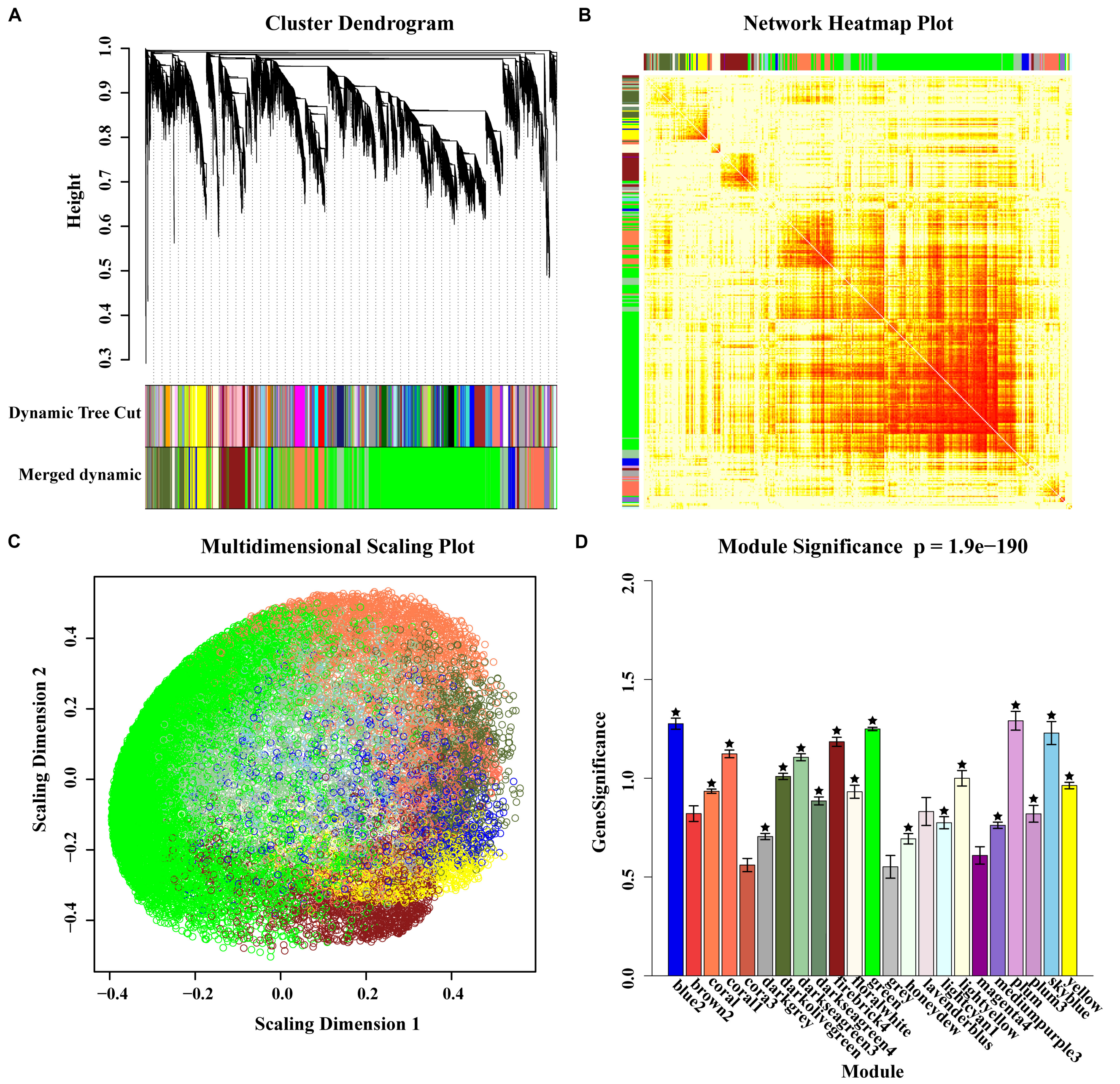


FIGURE 3 | Gene modules identified by weighted gene co-expression network analysis (WGCNA). **(A)** Gene dendrogram obtained by clustering the dissimilarity based on consensus Topological Overlap with the corresponding module colors indicated by the color row. Each colored row represents a color-coded module which contains a group of highly connected genes. A total of 23 modules were identified. **(B)** Heatmap plot of topological overlap in the gene network. In the heatmap, each row and column corresponds to a gene, light color denotes low topological overlap, and progressively darker red denotes higher topological overlap. Darker squares along the diagonal correspond to modules. The gene dendrogram and module assignment are shown along the left and top. **(C)** A multi-dimensional scaling plot of genes indicates that the green module is the most distinct. **(D)** Bar plot of mean gene significance across modules, and the star tagged module represents where the genes of selected regions are distributed.

Zm00001d028218 encodes a methionine-tRNA ligase, which catalyzes a reversible chemical reaction [from ATP, L-methionine and tRNA (Met) to AMP, diphosphate and L-methionyl-tRNA (Met)] in cytosol; and *Zm00001d028219* encodes a microtubule-associated protein (MAP70-2), which is closely related to MAP70-1 (Calder, 2010; Korolev et al., 2010; Oda and Fukuda, 2012). These results further indicate that the four genes we observed in our study are likely to play roles in the regulation of EL and FL, particularly in relation to the complex genetic mechanisms of EL and FL.

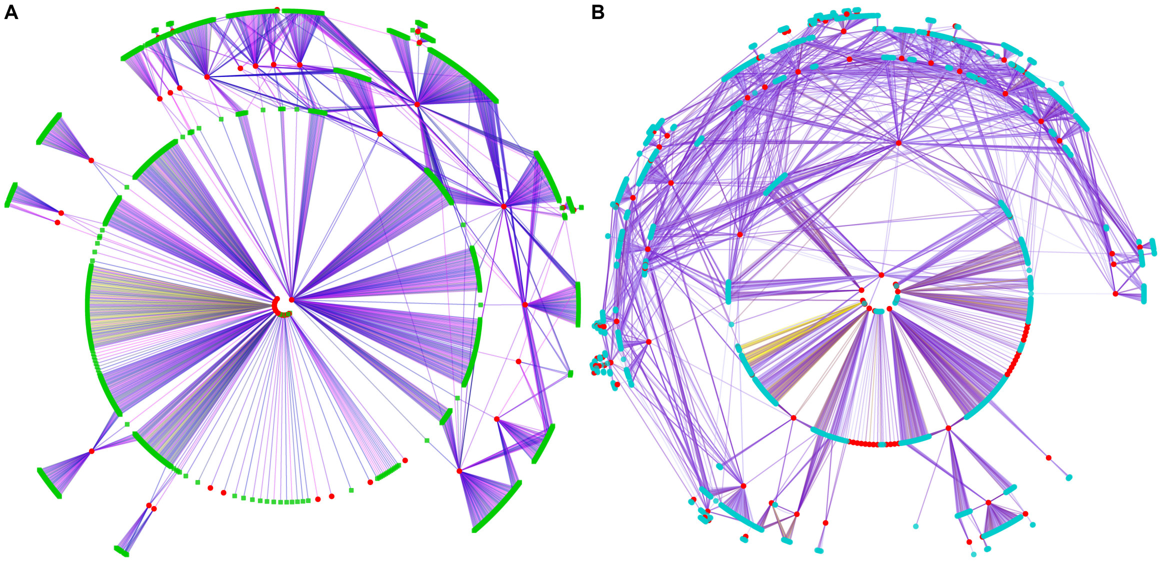


FIGURE 4 | Co-expression network of genes in selected regions of A54 and B154. Graphical view of the co-expression network where the nodes correspond to genes and the edges to co-expression links. Different colored edges represent interaction modules of different selection genes. **(A)** Co-expression network of selected genes of A54, and selected genes marked in red, interactive genes were marked in green. **(B)** Co-expression network of selected genes of B154, and selected genes marked in red, interactive genes were marked in blue.

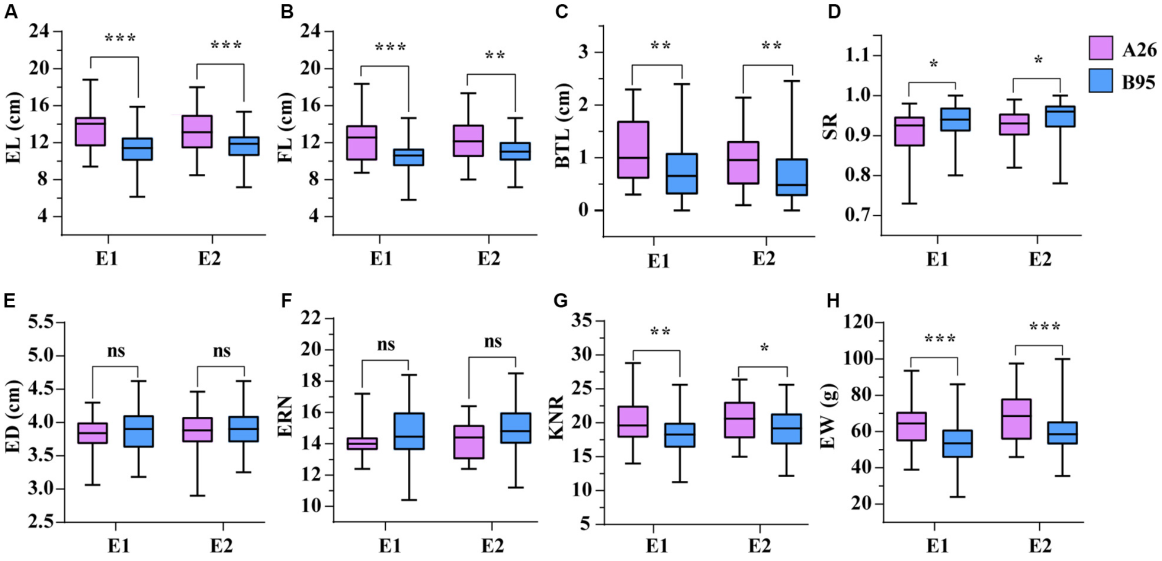


FIGURE 5 | The comparison of eight traits between A26 and B95 in two environments. **(A)** Ear length (EL, cm). **(B)** Fruit length (FL, cm). **(C)** Barren tip length (BTL, cm). **(D)** Setting rate (SR). **(E)** Ear diameter (ED, cm). **(F)** Ear row number (ERN). **(G)** Kernel number per row (KNR). **(H)** Ear weight (EW, g). *, **, and *** indicate significant level at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively. ns represents no significant difference.

Temporal and Spatial Expression Patterns and a Co-expression Network of Key Genes Associated With EL and FL

To explore and characterize the expression patterns of four target genes of EL and FL in a complex tissue, public RNA-seq datasets were used to analyze temporal and spatial expression variations of these genes in a large and diverse group of maize B73 tissues. Interestingly, *Zm00001d028216*, *Zm00001d028217*, and *Zm00001d028219* were expressed in specific tissues or developmental stages. However, the expression of *Zm00001d028218* was not remarkably different between any developmental stages, yet its expression level was higher than the three other genes in almost all tissues (**Figure 7**). *Zm00001d028216* and *Zm00001d028217* exhibited synchronous expression patterns in some specific

TABLE 1 | The basic statistics of ear length (EL), fruit length (FL), barren tip length (BTL) and setting rate (SR), ear weight (EW), ear diameter (ED), ear row number (ERN), and kernel number per row (KNR) of AM121 in two environments.

Trait	Environment	Range	Mean ± SD	SV (%)	H^2 (%)
EL (cm)	E1	6.13–18.80	11.79 ± 2.03	17.22	67.50
	E2	7.18–18.00	11.95 ± 1.88	15.73	
FL (cm)	E1	5.80–18.36	10.92 ± 1.92	17.58	69.84
	E2	7.18–17.34	11.23 ± 1.81	16.12	
BTL (cm)	E1	0.00–2.40	0.83 ± 0.58	69.88	65.33
	E2	0.00–2.46	0.71 ± 0.53	74.65	
SR	E1	0.73–1.00	0.93 ± 0.05	5.38	65.60
	E2	0.78–1.00	0.94 ± 0.04	4.26	
ED (cm)	E1	3.06–4.62	3.86 ± 0.32	8.29	75.18
	E2	2.90–4.62	3.90 ± 0.30	7.69	
ERN	E1	10.40–18.40	14.48 ± 1.61	11.12	66.75
	E2	11.20–18.50	14.82 ± 1.47	9.92	
KNR	E1	11.25–28.80	18.46 ± 3.25	17.61	69.39
	E2	12.20–26.40	19.36 ± 3.20	16.53	
EW (g)	E1	24.00–93.50	55.34 ± 12.43	22.46	79.09
	E2	35.50–100.00	61.07 ± 11.99	19.63	

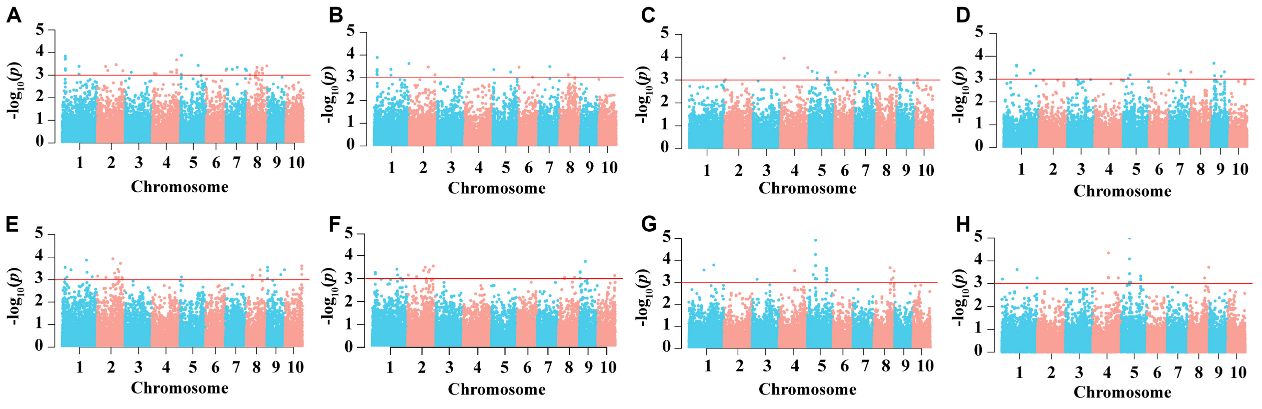


FIGURE 6 | Result of genome-wide association study. Manhattan plots of EL in E1 **(A)** and in E2 **(E)**, FL **(B)** and in E2 **(F)**, BTL **(C)** and in E2 **(G)** and SR in E1 **(D)** and in E2 **(H)**.

TABLE 2 | Co-localized regions detected by combining selective sweeps and GWAS.

Chr	Selected region	$\log_{10}$ (π ratio)	*Fst*	Associated region	Trait	Environment	*P*-value
1	26700001–26800000	0.902	0.301	26738827–27038887	EL, FL	E1, E2	5.16E-04
1	26800001–26900000	0.592	0.254	26738827–27038887	EL, FL	E1, E2	5.16E-04
1	303500001–303600000	0.315	0.512	303442010–303742023	FL	E1	2.36E-04
1	78400001–78500000	1.098	0.469	78287346–78587346	EL	E2	3.66E-04
1	78500001–78600000	1.179	0.491	78287346–78587346	EL	E2	3.66E-04
4	220900001–221000000	0.260	0.278	220783246–221083246	SR	E2	5.57E-04
4	238600001–238700000	0.387	0.285	238349228–238649228	BTL	E1	2.88E-04
4	32800001–32900000	0.710	0.498	32741698–33041698	BTL	E1	1.10E-04
6	121600001–121700000	0.648	0.449	121310639–121610639	FL	E2	9.37E-04
7	36700001–36800000	-1.132	0.328	36536622–36836622	BTL, SR	E1	6.12E-04
9	24000001–24100000	0.468	0.313	23836153–24136153	BTL	E1	8.00E-04
10	147000001–147100000	0.544	0.270	146955309–147393999	EL, FL	E2	7.16E-04
10	147100001–147200000	0.535	0.355	146955309–147393999	EL, FL	E2	7.16E-04

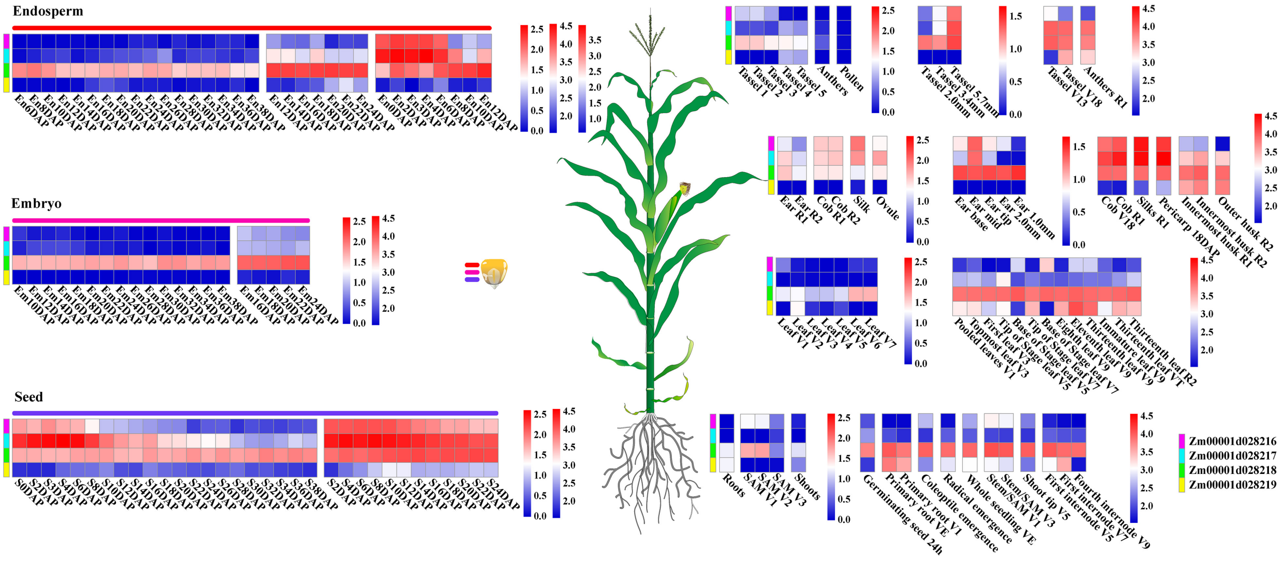


FIGURE 7 | Temporal and spatial expression patterns of regulation genes of EL and FL. Different expression modules in same tissue represent different expression data sources, and the scale bar shows the normalized expression values corresponded to different modules. Additionally, different colored bars represent four different genes.

developmental stages or tissues. For example, they were highly expressed in the early developmental stages of endosperm, seed, mid-ear, tassel (5.7 mm), silk, and pericarp, and especially in the cob. In particular, both *Zm00001d028216* and *Zm00001d028217* showed higher expression levels in the specific developmental stages of parts of the ear, silk, and ovule (**Figure 7**). These expression results indicate that *Zm00001d028216* and *Zm00001d028217* are likely to be key regulators during the determination of the meristem and the later differentiation and development of the ear. The synchronization and similarity in the expression characteristics of these genes also suggest that they coordinate and regulate the development of certain tissues, but the hub genes between them remain unknown.

To gain further insight into the relationship between these genes, and the influence of external genes, genes that potentially interact with them were extracted from the co-expression network to construct sub-networks (based on a weight value greater than 0.2). The most important finding of the co-expression network was that some hub genes were involved in sub-networks that had *Zm00001d028216* and *Zm00001d028217* at the core (**Figure 8**). These hub genes were composed of 28 protein-coding genes, and included three transcription factors. They were mainly involved in cell differentiation, cellular developmental processes, and anatomical structure morphogenesis (**Supplementary Table S8**). Nevertheless, the other two genes constituted a relatively independent network, and they were functionally enriched in a number of critical developmental processes. This indicates that EL and FL are affected by diverse biological processes, and multilevel genes may be involved in communicating and adjusting the function balance among these hub genes.

DISCUSSION

During crop improvement, phenotypic characteristic developments are dependent on market demand and breeder selection. Micro-evolution during the breeding process not only increases genetic divergence, but also changes genetic diversity. In this study, 32,306 high-quality SNPs were used to identify the effect of artificial improvement across the genome, and our analysis suggests that there has been genetic divergence between the Shaan A and B groups in AM208. These results indicate that artificial selection was effective in causing population divergence between Shaan A and Shaan B heterotic groups. This is consistent with previous findings that selection in breeding programs makes significant contributions to genetic divergence between heterotic groups (Liu et al., 2003). Moreover, B154 is more dispersed and has a shorter LD decay distance than A54. Also, the median π ratio (πA54/πB154 is equal to 0.854) further supports these differences. Our results also indicate that B154 contains a richer genetic diversity than A54. Additionally, it is particularly interesting that B154, which has a richer genetic diversity, has more selection regions. Our analysis provides new and important information on the breeding history genomics of Shaan A and Shaan B groups.

Furthermore, a subset of 337 regions including 772 genes in the genome that are related to artificial selection were identified. Within the 772 selected genes, approximately 24 candidate genes including *Zm00001d028217* have been reported in a previous study to be associated with modern breeding selection (Jiao et al., 2012). In addition, 20 including *Zm00001d028216* and 22 candidate genes overlapped with candidate genes related to maize domestication and improvement, respectively (Hufford et al., 2012; **Supplementary Table S9**). These combined results

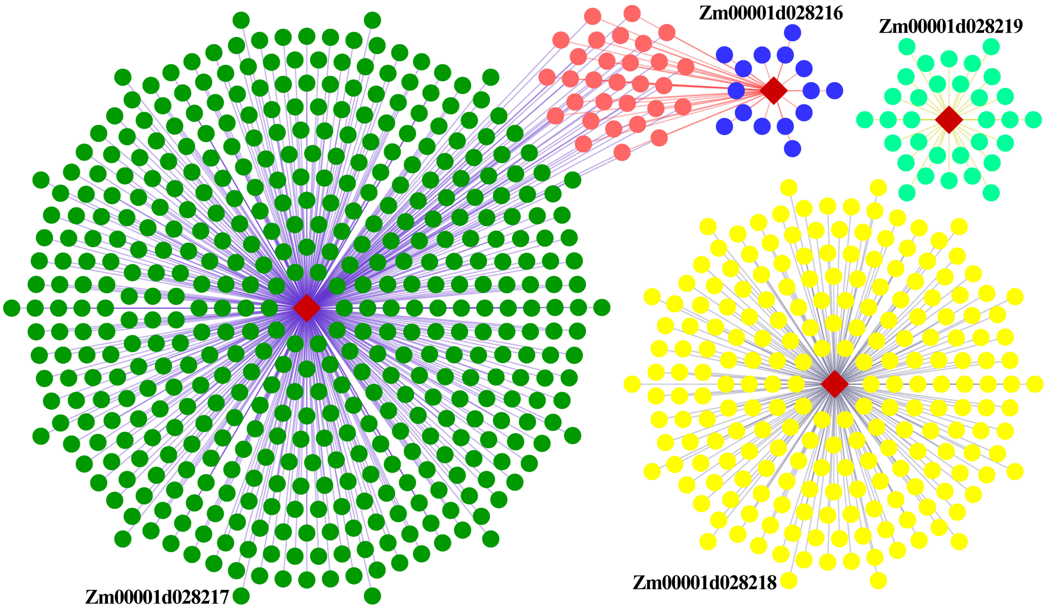


FIGURE 8 | Regulatory co-expression network of key genes related to EL and FL. Weighted gene co-expression gene network of multiple maize tissues, visualized using an organic layout in cytoscape. The nodes in the network were obtained by coloring each gene by its specific expression profile along maize tissues development. Different genes and their interactive genes were marked with different color. Of which, blue nodes represent weight value more than 0.5, orange nodes represent common interactive genes of *Zm00001d028216* and *Zm00001d028217*.

indicate that these genes are of high value for crop improvement. A large proportion of identified genes have not been reported because they are from different populations, are related to different focus traits from different breeders, or are identified using different genotyping technology. We found that there are distinctive differences in gene numbers in selected regions between A54 and B154. Enrichment analysis revealed that these genes are involved in regulating different essential metabolic processes during maize development (due to the different usage of their materials during the breeding process). Moreover, the co-expression networks captured important biological modules of selected genes between A54 and B154. These modules showed that these genes were not likely to be independently regulating these metabolic processes, and were also not the only key factors in causing the divergence in some target traits. Additionally, the selected genes and the interactive genes were only a little enriched in response to abiotic stress. This is likely to be because these inbreds were selected under the same conditions: low irrigation rates, low nitrogen levels, and high plant densities. The interactive genes with selected genes were enriched in the regulation of the timing of meristematic phase transition and transition from the vegetative to reproductive phase, as well as pollen germination. These are important targets for a female and male plant, implying that genotype and phenotype difference may be generate gradually during improvement of heterosis groups. Taken together, our results indicate that artificial selection alters consciously or unconsciously the alleles' frequencies of target traits during improvement process.

In the maize breeding process, construction of heterosis groups is considered to be the most important element. Hybrid vigor remarkably contributes to the performance of hybrids compared to their parents (Birchler et al., 2003). Therefore, parents' inbred lines may have obvious complementation in the trait and genome. EL, FL, BTL, and SR are vital to increasing yield. In this study, A54 performed longer EL and FL than B154. However, B154 showed higher SR (equivalent to shorter BTL) than A54. The combination crossing from these two groups may produce hybrids with long ears and high SR, such as the SD650 (KA105/KB024) and SD636 (KA103/KB043) hybrids that have been widely grown in Shaanxi province. In this study, four candidate genes related to EL and FL were identified by combining GWAS and a selective sweep. Among the four genes, *Zm00001d028216* and *Zm00001d028217* also were identified as selected genes in maize domestication and improvement processes (Hufford et al., 2012; Jiao et al., 2012). Moreover, *Zm00001d028217* belongs to the MADS family and *Zm00001d028216* belongs to the YABBY family, these two transcription factors play important roles in regulating spikelet development, floral induction, and inflorescence development (Cacharr et al., 1999; Laudenciachingcuanco and Hake, 2002; Danilevskaya et al., 2008; Thompson and Hake, 2009). Alternatively, *Zm00001d028216* and *Zm00001d028217* showed higher expression levels at the development stage and are known to regulate other genes involved in cell differentiation, cellular developmental processes, and anatomical structure morphogenesis. Therefore, it is meaningful and interesting to carry out further validation work on these four genes to understand the underlying molecular biology mechanism. In addition, short BTL (higher SR) is an desirable trait for high-yield breeding, however, genetic basis of BTL and SR remain poorly understand (Li et al., 2020). In the present study, only five associated regions related to SR and BTL overlapped with the selected regions. Moreover, many significant SNPs associated with SR and BTL were identified through association analysis. Therefore, this study has helped to uncover the genetic mechanism of SR and BTL, information which can be used to aid genetic improvements. Nonetheless, some associated regions related to SR and BTL were detected, and more research is needed to identify the functional genes, particularly in the overlapped region.

Genetic analysis for complex traits generally requires genetic populations with a diverse phenotype. This may lead to a situation where researchers collect diverse materials from multiple breeding programs and geographical areas but do not focus on studying breeding materials from a single origin. We used 208 inbred lines selected from the Shaan A and Shaan B groups that were derived from a common ancestral pool but used different testers. As breeding populations, Shaan A and Shaan B groups integrate a diverse germplasm and allele frequency which have been subjected to noteworthy changes due to long-term artificial selection. More importantly, as homozygotes, inbred lines from Shaan A and Shaan B groups can be used as ideal test materials for genetic studies and marker-assisted breeding. In future, genetic studies that use breeding populations will greatly progress breeding applications. Our study used these inbreds from a breeding population to investigate genetic divergence and we identified the selective regions related to traits. This not only provides insights into the effects of artificial selection across the genome for crop improvement but also conveys essential information on some important agronomic traits.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

SX and JX conceived and designed the experiments. XZ, YH, and XT contributed to field management. YW, YL, and NW contributed to the collection of phenotypic data and DNA extraction. TL and JQ analyzed the data and wrote the manuscript. SX contributed to revising the manuscripts. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Innovation Project of National Key R&D Program of China (Grant No. 2018YFD0100200), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 31701438), and the Shaanxi Province Key Research and Development Program of China (2017ZDCXL-NY-02-04). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

ACKNOWLEDGMENTS

We are grateful to Prof. Jianming Yu and Prof. Qin Yang for revising the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00747/full#supplementary-material

FIGURE S1 | GO analysis of interactive genes with selected genes in A54 and B154. Histogram showing the significantly functional distribution of interactive genes with selected genes in A54 and B154, respectively. The scale bar represents the significant levels.

FIGURE S2 | Data distribution and correlation analysis of eight traits. *, **, and *** indicate a significant correlation at $P < 0.05$, $P < 0.01$, and $P < 0.001$ respectively.

FIGURE S3 | QQ-plots of eight ear related traits in two locations. QQ-plot of EL in E1 **(A)** and E2 **(E)**. QQ-plot of FL in E1 **(B)** and E2 **(F)**. QQ-plot of BTL in E1 **(C)** and E2 **(G)**. QQ-plot of SR in E1 **(D)** and E2 **(H)**.

TABLE S1 | The material information of AM208.

TABLE S2 | The list of selective regions and genes in A54 and B154, respectively.

TABLE S3 | GO enrichment analysis of selected genes from Shaan A and B groups.

TABLE S4 | Interaction between selected genes and other genes.

TABLE S5 | GO enrichment analysis of interactive genes of selected genes.

TABLE S6 | Results of genome-wide association study for EL, FL, BTL, and SR in two environments.

TABLE S7 | Haplotype effect of three significant SNPs related to EL and FL.

TABLE S8 | Functional enrichment analysis of key genes regulating EL and FL.

TABLE S9 | Common selected genes in maize domestication and improvement processes compared to previous studies.

REFERENCES

Assenov, Y., Ramírez, F., Schelhorn, S. E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics* 24, 282–284. doi: 10.1093/bioinformatics/btm554

Birchler, J. A., Auger, D. L., and Riddle, N. C. (2003). In search of the molecular basis of heterosis. *Plant Cell* 15, 2236–2239. doi: 10.1105/tpc.151030

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Browning, B., and Browning, S. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Cacharr, N. J., Saedler, H., and Theissen, G. (1999). Expression of MADS box genes ZMM8 and ZMM14 during inflorescence development of Zea mays discriminates between the upper and the lower floret of each spikelet. *Dev. Genes Evol.* 209, 411–420. doi: 10.1007/s004270050271

Calder, G. (2010). The microtubule-associated protein AtMAP70-5 regulates secondary wall patterning in *Arabidopsis* wood cells. *Curr. Biol.* 20, 744–749. doi: 10.1016/j.cub.2010.02.057

Chen, H., and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12:35. doi: 10.1186/1471-2105-12-35

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., et al. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* 166, 252–264. doi: 10.1104/pp.114.240689

Chen, Q., Han, Y., Liu, H., Wang, X., Sun, J., Zhao, B., et al. (2018). Genome-wide association analyses reveal the importance of alternative splicing in diversifying gene function and regulating phenotypic variation in maize. *Plant Cell* 30, 1404–1423. doi: 10.1105/tpc.18.00109

Chong, X., Fei, Z., Wu, Y., Yang, X., and Chen, F. (2017). A SNP-enabled assessment of genetic diversity, evolutionary relationships and the identification of candidate genes in chrysanthemum. *Genome Biol. Evol.* 8, 3661–3671.

Danilevskaya, O. N., Meng, X., Selinger, D. A., Deschamps, S., Hermon, P., Vansant, G., et al. (2008). Involvement of the MADS-box gene ZMM4 in floral induction and inflorescence development in maize. *Plant Physiol.* 147, 2054–2069. doi: 10.1104/pp.107.115261

Doebley, J., Stec, A., and Gustus, C. (1995). teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141, 333–346.

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell* 127, 1309–1321. doi: 10.1016/j.cell.2006.12.006

Hallauer, A. R., and Miranda, F. J. B. (1981). Quantitative genetics in maize breeding. *Q. Rev. Biol.* 6, 124–126.

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D. A., Yang, Z., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytol.* 209:871. doi: 10.1111/nph.13626

Heuer, S., Hansen, S., Bantin, J., Brettschneider, R., Kranz, E., Lörz, H., et al. (2001). The maize MADS box gene ZmMADS3 affects node number and spikelet development and is co-expressed with ZmMADS1 during flower development, in egg cells, and early embryogenesis. *Plant Physiol.* 127, 33–45. doi: 10.1104/pp.127.1.33

Huang, J., Vendramin, A. S., Shi, L., and Mcginnis, K. (2017). Construction and optimization of large gene co-expression network in maize using RNA-Seq data. *Plant Physiol.* 175, 568–583. doi: 10.1104/pp.17.00825

Hufford, M. B., Xu, X., Heerwaarden, J. V., Pyhäjärvi, T., Chia, J. M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44, 808–811.

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., et al. (2012). Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* 46, 812–815. doi: 10.1038/ng.2312

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Korolev, A. V., Chan, J., Naldrett, M. J., Doonan, J. H., and Lloyd, C. W. (2010). Identification of a novel family of 70&emsp;kDa microtubule-associated proteins in *Arabidopsis* cells. *Plant J.* 42, 547–555. doi: 10.1111/j.1365-313x.2005.02393.x

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Laudenciachingcuanco, D., and Hake, S. (2002). The indeterminate floral apex1 gene regulates meristem determinacy and identity in the maize inflorescence. *Development* 129, 2629–2638.

Li, G., Wang, D., Yang, R., Logan, K., Chen, H., Zhang, S., et al. (2014). Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, 7582–7587. doi: 10.1073/pnas.1406383111

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, T., Qu, J., Wang, Y., Chang, L., He, K., Guo, D., et al. (2018). Genetic characterization of inbred lines from Shaan A and B groups for identifying loci associated with maize grain yield. *BMC Genet.* 19:63. doi: 10.1186/s12863-018-0669-9

Li, Z., Liu, P., Zhang, X., Zhang, Y., Ma, L., Liu, M., et al. (2020). Genome-wide association studies and QTL mapping uncover the genetic architecture of ear tip-barrenness in maize. *Physiol. Plant.* [Epub ahead of print]. doi: 10.1111/ppl.13087

Liu, K., Goodman, M., Muse, S., Smith, J. S., Buckler, E., and Doebley, J. (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165, 2117–2128.

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325.

Oda, Y., and Fukuda, H. (2012). Secondary cell wall patterning during xylem differentiation. *Curr. Opin. Plant Biol.* 15, 38–44. doi: 10.1016/j.pbi.2011.10.005

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Sarkar, N. K., Kim, Y. K., and Grover, A. (2014). Coexpression network analysis associated with call of rice seedlings for encountering heat stress. *Plant Mol. Biol.* 84, 125–143. doi: 10.1007/s11103-013-0123-3

Schmutz, J., Mcclean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Setter, T. L., Flannigan, B. A., and Melkonian, J. (2001). Loss of kernel set due to water deficit and shade in maize. *Crop Science* 41, 1530–1540. doi: 10.2135/cropsci2001.4151530x

Shi, J., and Lai, J. (2015). Patterns of genomic changes with crop domestication and breeding. *Curr. Opin. Plant Biol.* 24, 47–53. doi: 10.1016/j.pbi.2015.01.008

Souza, C. L. D., Barrios, S. C. L., and Moro, G. V. (2010). Performance of maize single-crosses developed from populations improved by a modified reciprocal recurrent selection. *Sci. Agric.* 67, 198–205. doi: 10.1590/s0103-90162010000200011

Thompson, B. E., and Hake, S. (2009). Bearded-ear encodes a MADS box transcription factor critical for maize floral development. *Plant Cell* 21:2578. doi: 10.1105/tpc.109.067751

Van Heerwaarden, J., Hufford, M. B., and Ross-Ibarra, J. (2012). Historical genomics of North American maize. *Proc. Natl. Acad. Sci. U.S.A.* 109, 12420–12425. doi: 10.1073/pnas.1209275109

Viana, J. M. S., DeLima, R. O., Mundim, G. B., Cond, A. B. T., and Vilarinho, A. A. (2013). Relative efficiency of the genotypic value and combining ability effects on reciprocal recurrent selection. *Theor. Appl. Genet.* 126, 889–899. doi: 10.1007/s00122-012-2023-3

Wang, R., Stec, A., Hey, J., Lukens, L., and Doebley, J. (1999). The limits of selection during maize domestication. *Nature* 398, 236–239. doi: 10.1038/18435

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., and Rokas, A. (2017). A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29, 944–959. doi: 10.1105/tpc.17.00009

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide association studies in maize: praise and stargaze. *Mol. Plant* 10, 359–374. doi: 10.1016/j.molp.2016.12.008

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, 316–322.

Xie, W., Wang, G., Yuan, M., Yao, W., and Zhang, Q. (2015). Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl. Acad. Sci. U.S.A.* 112:E5411.

Xu, G. (2018). *Genetic Basis of Artificial Selection Response in High-Oil Maize*. Beijing: China Agricultural University.

Yamasaki, M., Tenaillon, M. I., Bi, I. V., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., et al. (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* 17:2859. doi: 10.1105/tpc.105.037242

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yu, G., Wang, L.-G., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, H., Jiao, B., and Liang, C. (2017). Systematic analysis Of RNA-Seq-based gene co-expression across multiple plants. *bioRxiv* [Preprint]. doi: 10.1101/139923

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhang, Z., Li, H., Zhang, D., Liu, Y., Jing, F., Shi, Y., et al. (2012). Characterization and expression analysis of six MADS-box genes in maize (*Zea mays* L.). *J. Plant Physiol.* 169, 797–806. doi: 10.1016/j.jplph.2011.12.020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fgene.2020.01021

Transcriptome Sequencing and Development of Novel Genic SSR Markers From *Pistacia vera* L.

Harun Karcı[1], Aibibula Paizila[1], Hayat Topçu[1], Ertuğrul Ilikçioğlu[2] and Salih Kafkas[1]*

[1] Department of Horticulture, Faculty of Agriculture, Çukurova University, Adana, Turkey, [2] Pistachio Research Institute, Gaziantep, Turkey

In this study, we aimed to develop novel genic simple sequence repeat (eSSR) markers and to study phylogenetic relationship among *Pistacia* species. Transcriptome sequencing was performed in different tissues of Siirt and Atlı cultivars of pistachio (*Pistacia vera*). A total of 37.5-Gb data were used in the assembly. The number of total contigs and unigenes was calculated as 98,831, and the length of N50 was 1,333 bp after assembly. A total of 14,308 dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs (4–17) were detected, and the most abundant SSR repeat types were trinucleotide (29.54%), dinucleotide (24.06%), hexanucleotide (20.67%), pentanucleotide (18.88%), and tetranucleotide (6.85%), respectively. Overall 250 primer pairs were designed randomly and tested in eight *Pistacia* species for amplification. Of them, 233 were generated polymerase chain reaction products in at least one of the *Pistacia* species. A total of 55 primer pairs that had amplifications in all tested *Pistacia* species were used to characterize 11 *P. vera* cultivars and 78 wild *Pistacia* genotypes belonging to nine *Pistacia* species (*P. khinjuk*, *P. eurycarpa*, *P. atlantica*, *P. mutica*, *P. integerrima*, *P. chinensis*, *P. terebinthus*, *P. palaestina*, and *P. lentiscus*). A total of 434 alleles were generated from 55 polymorphic eSSR loci with an average of 7.89 alleles per locus. The mean number of effective allele was 3.40 per locus. Polymorphism information content was 0.61, whereas observed (Ho) and expected heterozygosity (He) values were 0.39 and 0.65, respectively. UPGMA (unweighted pair-group method with arithmetic averages) and STRUCTURE analysis divided 89 *Pistacia* genotypes into seven populations. The closest species to *P. vera* was *P. khinjuk*. *P. eurycarpa* was closer *P. atlantica* than *P. khinjuk*. *P. atlantica*–*P. mutica* and *P. terebinthus*–*P. palaestina* pairs of species were not clearly separated from each other, and they were suggested as the same species. The present study demonstrated that eSSR markers can be used in the characterization and phylogenetic analysis of *Pistacia* species and cultivars, as well as genetic linkage mapping and QTL (quantitative trait locus) analysis.

Keywords: transcriptome, RNA-seq, eSSR, *Pistacia*, phylogenetic

OPEN ACCESS

Edited by:
Yin Li,
Rutgers, The State University of New Jersey, United States

Reviewed by:
Sezai Ercisli,
Atatürk University, Turkey
Murat Guney,
Bozok University, Turkey

****Correspondence:***
Salih Kafkas
skafkas@cu.edu.tr

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 June 2020*
Accepted: *10 August 2020*
Published: *09 September 2020*

Citation:
Karcı H, Paizila A, Topçu H, Ilikçioğlu E and Kafkas S (2020) Transcriptome Sequencing and Development of Novel Genic SSR Markers From Pistacia vera L.. Front. Genet. 11:1021. doi: 10.3389/fgene.2020.01021

INTRODUCTION

Pistacia L. genus is a member of the Anacardiaceae family that also contains important species such as mango, pepper tree, and sumac (Kafkas, 2006a). The sex habit of *Pistacia* is dioecious with several exceptions (Kafkas et al., 2000). The genus of *Pistacia* consists of 13 or more species (Gundesli et al., 2019), and *Pistacia vera* is believed to be the most ancestral species, whereas the other species probably derived from Zohary (1952) and Kafkas (2019). Currently, the main pistachio producers in the world are Iran, the United States, Turkey, and Syria (FAO, 2019).

The first taxonomic study in the genus *Pistacia* was done morphologically by Zohary (1952). After discovery of different molecular markers, they have been used in *Pistacia* as well. The first detailed molecular study in *Pistacia* was performed based on chloroplast DNA profiles by Parfitt and Badenes (1997). Microsatellites or simple sequence repeats (SSRs) and repeats of 1- to 6-nucleotide-long DNA motifs have high reproducibility, multiallelic character, and extensive tandem repeats in the whole genome (Powell et al., 1996). SSRs have advantages over other marker systems because of their codominant inheritance, suitability for automation, and well-distribution throughout eukaryotic genomes. Recently, SSRs have been widely used in genetic map construction, DNA fingerprinting, genetic diversity, quantitative trait locus (QTL) mapping, and marker-assisted selection (MAS) (Dong et al., 2018; Yang et al., 2018; Zhang et al., 2019).

Genic SSRs or eSSRs are obtained by expression sequence tags that are created by gene transcripts that have been converted into cDNA (Adams et al., 1991). Recently, eSSR markers have been used for identifying plant species because of its design from coding regions (Varshney et al., 2005; Ellis and Burke, 2007). The concern here is that because eSSRs are located within the genes, and more conserved, they may be used for identification of alleles related to some agronomically significant traits (Chen et al., 2017; Dong et al., 2018; Yang et al., 2018). Conventionally, SSR development needs to be labor-intensive, such as cloning DNA and constructing library, and generates particularly less SSRs compared with next-generation sequencing (NGS) technology (Zalapa et al., 2012; Zhang et al., 2012). The advantage of NGS technologies, especially next-generation transcriptome sequencing, provides a large amount of sequences with cost-effective and high-quality data in a short period (Wang et al., 2010; Taheri et al., 2019; Zhang et al., 2019).

RNA sequencing (RNA-seq) is considered to provide information about functional genes, used to detect reliable and high-throughput eSSR markers (Wang et al., 2018; Taheri et al., 2019). Using RNA-seq, eSSRs have been reported in several plant species such as bean (Chen et al., 2015), grape (Huang et al., 2011), coffee (Ferrão et al., 2015), tomato (Zhou et al., 2015), barley (Zhang et al., 2014), cotton (Tabbasam et al., 2014), wheat (Gupta et al., 2003; Gadaleta et al., 2009), cucumber (Hu et al., 2010), walnut (Zhu et al., 2009), and citrus (Chen et al., 2006).

The aims of the study were (i) to develop novel genic SSR markers from transcriptome sequences of pistachio (*P. vera*) and (ii) to determine phylogenetic relationship among *Pistacia* species using novel eSSRs.

MATERIALS AND METHODS

Plants Materials

In this study, RNA isolation and transcriptome sequencing were performed in bud, flower, leaf, shoot, whole nut, pericarp, and kernel of Siirt (female) and Atlı (male) cultivars (**Supplementary Table 1**). The sampled tissues were immediately frozen in liquid nitrogen and stored at -80°C until RNA isolation.

The genomic DNAs belonging to 11 *P. vera* L., 5 *P. khinjuk* stocks, 13 *P. atlantica* Desf., 5 *P. mutica* F.&M., 3 *P. atlantica* × *P. integerrima* hybrids (UCB1), 8 *P. integerrima* Stewart, 8 *P. chinensis* Bunge, 11 *P. terebinthus* L., 5 *P. palaestina* Boiss., 12 *P. lentiscus* L., and 6 unknown *Pistacia* genotypes were used to verify eSSR markers. The leaf samples from these genotypes were collected from germplasm collections in Çukurova University in Adana, Pistachio Research Institute in Gaziantep. Wild *P. atlantica* and *P. eurycarpa* genotypes growing in the nature in Manisa and Gaziantep provinces were used as well. Two *P. eurycarpa* genotypes were about 500–600 years old in Göbek village (**Figure 1** and **Supplementary Table 2**).

RNA Extraction, Sequencing, Transcriptome Assembly

Total RNA was extracted from different tissues of pistachio and was sequenced by BGI (Beijing Genomic Institute) using an Illumina (Hi-Seq 2500) NGS platform. The raw reads were first cleaned from adaptors and filtered for low-quality sequences with higher than 20% Q value < 20 bases. Those called as "clean reads" were assembled using Trinity software (v2.0.6) (Grabherr et al., 2011). After *de novo* assembly, Trinity sequences were named as transcripts obtained from contigs that were not extended on each ends The transcripts were clustered and obtained final unigenes with TGIGL (Pertea et al., 2003) software. Simple sequence repeats were searched using MIcroSAtelite (MISA) (Haas et al., 2013) search module in all unigenes. The search parameters were set for the detection of dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs with a minimum of six, five, four, four, and four repeats, respectively. Primers pairs were designed using online software Batchprimer 3.0 with the standard parameters (You et al., 2008) (**Supplementary Table 3**). The sequence data have been deposited in the National Center for Biotechnology Information under BioProject accession number PRJNA648340.

DNA Extraction, Polymerase Chain Reaction Amplification, eSSR Validation

The genomic DNAs of 89 *Pistacia* samples were extracted from fresh leaves using the CTAB method of Doyle and Doyle (1987) with minor modifications (Kafkas et al., 2006). DNA concentrations were measured using a Qubit Fluorimeter (Invitrogen) or were estimated by comparing the band intensity with λ DNA of known concentrations following 0.8% agarose gel electrophoresis and ethidium bromide staining. DNA samples were subsequently diluted to a concentration of 10 ng/μL for eSSR–polymerase chain reaction (PCR).

FIGURE 1 | Pictures of two old *P. eurycarpa* genotypes. **(A)** PE-1, **(B)** PE-2.

Initially, a total of 250 randomly selected primer pairs were screened in eight *Pistacia* individuals, and 55 primer pairs were selected for the characterization of 89 *Pistacia* genotypes belonging to 10 *Pistacia* species.

eSSR-PCR was carried out using a three-primer strategy according to the method utilized by Schuelke (2000) with minor modifications (Topçu et al., 2016). PCR was performed in a total volume of 12.5 μL containing 20 ng DNA; 75 mM Tris–HCl (pH 8.8); 20 mM $(NH_4)_2SO_4$; 2.0 mM $MgCl_2$; 0.01% Tween 20; 200 μM each dNTP; 10 nM M13 tailed forward primer at the 5′ end; 200 nM reverse primer; 200 nM universal M13 tail primer (5 TGTAAAACGACGGCCAGT-3) labeled with one of FAM, VIC, NED, or PET dyes; and 0.6 U hot-start Taq DNA polymerase. Amplification was performed in two steps as follows: initial denaturation at 94°C for 3 min, followed by 10 cycles at 94°C for 30 s, 58°C for 45 s, and 72°C for 60 s. The second step included 30 cycles at 94°C for 30 s, 58°C for 45 s, and 72°C for 60 s and a final extension at 72°C for 10 min. After the PCR was completed, the reactions were subjected to denaturation for capillary electrophoresis in an ABI 3130xl genetic analyzer [Applied Biosystems Inc., Foster City, CA, United States (ABI)] using a 36-cm capillary array with POP7 as the matrix (ABI). Samples were fully denatured by mixing 0.5 μL of the amplified product with 0.2 μL of the size standard and 9.8 μL formamide. The fragments were resolved using ABI data collection software 3.0, and SSR fragment analysis was performed with GeneScan Analysis Software 4.0 (ABI).

Genetic Diversity

The 55 polymorphic eSSR loci were used for the genetic diversity of 11 *P. vera* and 78 wild *Pistacia* genotypes. The number of alleles (Na), the number of effective alleles (Ne), observed (Ho), and expected (He) heterozygosity were calculated using GenAlEx version 6.5 according to Peakall and Smouse (2012). The polymorphism information contents (PICs) of each marker were calculated using PowerMarker software version 3.25 (Liu and Muse, 2005). The SSR bands scored as present (1) or absent (0) consisted of a dendrogram using NTSYSpc v2.21c (Rohlf, 2009) software by unweighted pair-group method with arithmetic averages (UPGMA). A principal coordinate analysis (PCoA) was performed using NTSYSpc v2.21c (Rohlf, 2009).

STRUCTURE 2.3.4 software (Pritchard et al., 2000) was also used to determine the possible number of populations and for the construction of the population structure. Structure analysis computes the proportion of the genome of an individual originating from each interfered population. Possible *K*'s (where *K* is an assumed fixed number of subpopulations in the entire population) were from 1 to 10 with five replications to ensure consistency of results. Ln P(D)s mean possible estimated *K*'s. There are Ln P(D) values for each *K* value. By using Ln P(D) values for every *K* calculate Δ*K* that shows a possible number of populations. Each replication run was conducted with a burn-in period of 100,000 and 100,000 Markov chain Monte Carlo.

RESULTS

Sequencing and Assembly

Sixteen transcriptome libraries were constructed from different tissues of Siirt (female) and Atlı (male) cultivars, and a total of 374,726,850 clean reads were obtained. A total of 98,831 unigenes were generated by the Trinity software, and the N50 of unigenes was computed as 1,333 bp (**Supplementary Table 4**). All unigenes were classified according to size of the sequences; 55,101 (55.75%) were 100–500 bp, 18,462 (18.68%) were 500 bp–1 kb, 10,615 (10.74%) were 1–1.5 kb, 6,804 (6.88%) were 1.5–2 kb, and the rest of sequences 7,849 (7.94%) were >2 kb (**Table 1**).

Identification and Distribution of SSR Motifs

The MISA search module was used to search for SSRs with the 98,831 unigenes. In total, 37,793 potential genic SSRs were identified from 98,831 unigene sequences, of which 23,485 were mononucleotide repeats (**Supplementary Table 5**).

A total of 14,308 dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs were detected, and the most abundant type of the repeats was

TABLE 1 | Length distribution of assembled unigenes.

Nucleotides length	Unigenes
100–500 bp	55,101
500–1 kb	18,462
1–1.5 kb	10,615
1.5–2 kb	6,804
>2 kb	7,849
N50 bp	1,333
Mean length (bp)	776
Max length (bp)	660
Min length (bp)	379
All unigenes	98,831

trinucleotide motifs (29.54%), followed by dinucleotide (24.06%), hexanucleotide (20.67%), pentanucleotide (18.88%), and tetranucleotide motifs (6.85%) (**Table 2** and **Supplementary Table 6**). The most abundant repeats were AG/CT (15.4%), AAG/CTT (9.8%), and AT/AT (6.5%). The most abundant tetranucleotide, pentanucleotide, and hexanucleotide repeat motif types were AAAT/ATTT (2.6%), AAAAT/ATTTT (4.7%), and AAAAAT/ATTTTT (2.4%), respectively (**Figure 2**).

Validation of SSRs

The 250 genic SSR primer pairs were designed from assembled short RNA sequences. The 250 eSSR primer pairs were screened in eight *Pistacia* species (*P. vera*, *P. khinjuk*, *P. atlantica*, *P. mutica*, *P. integerrima*, *P. chinensis*, *P. terebinthus*, and *P. lentiscus*) by agarose gel electrophoresis (1.5%). In total, 233 were amplified at least in one of the *Pistacia* species, and 82 were amplified in all tested eight *Pistacia* species (**Supplementary Table 7**). Of them, 55 eSSR primer pairs were used for characterization of 89 *Pistacia* accessions because of their polymorphism and having amplification in the tested *Pistacia* species.

Functional Annotation and Classification

A total of 98,831 assembled unigenes were aligned to different universal databases. Annotation demonstrated that 52,839 unigenes (53.46%) were found important in the Nr database, 49,727 (50.31%) in the Nt database, and 34,419 (34.82%) in the Swiss-Prot database. The annotation of 58,401 (59.09%) unigenes was successfully achieved in at least one of the six public databases (**Supplementary Table 8**).

According to Gene Ontology (GO) analysis, there were 40,405 annotated unigenes classified into three functional categories: biological process, cellular component, and molecular function. The largest classes in biological process were "cellular process" (25,241), "metabolic process" (24,781), and "single organism process" (17,507), respectively. The cellular component category mostly consists of proteins involved in "cell" (29,529), "cell part" (29,528), and "organelle" (23,825). The highest classes in the molecular function category were detected as "binding" (19,780), "catalytic activity" (20,554), and "transporter" (2,654), respectively (**Supplementary Table 9** and **Figure 3**).

The unigenes were aligned to COG database, and classification of 19,349 (19.58%) unigenes was divided into 25 specific categories (**Supplementary Table 10** and **Figure 4**). Among these categories, "general function prediction only" (6,665), "transcription" (3,584), "replication, recombination and repair" (3,276), "signal transduction mechanisms" (2,774), "posttranslational modification, protein turnover, chaperones" (2,627), and "translation, ribosomal structure, and biogenesis" (2,343) formed the largest groups.

KEGG Orthology (KO) database was used for metabolic pathway analysis. In the results constructed in a total 128 KEGG

TABLE 2 | Number of repeats; number of dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs; total number of SSR motifs; percentage of SSR motifs; and total number of SSR motifs in *P. vera*.

No. of repeats	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
4	0	0	0	2,286	2,467
5	0	0	736	361	486
6	0	2,354	201	54	4
7	1,228	1,139	43	1	0
8	714	590	0	0	0
9	449	136	0	0	0
10	414	3	0	0	0
11	345	3	0	0	0
12	238	1	0	0	0
13	51	0	0	0	0
14	1	0	0	0	0
15	2	0	0	0	0
16	0	0	0	0	0
17	1	0	0	0	0
Total	3,443	4,226	980	2,702	2,957
%	24.06	29.54	6.85	18.88	20.67
Total of repeats			14,308		

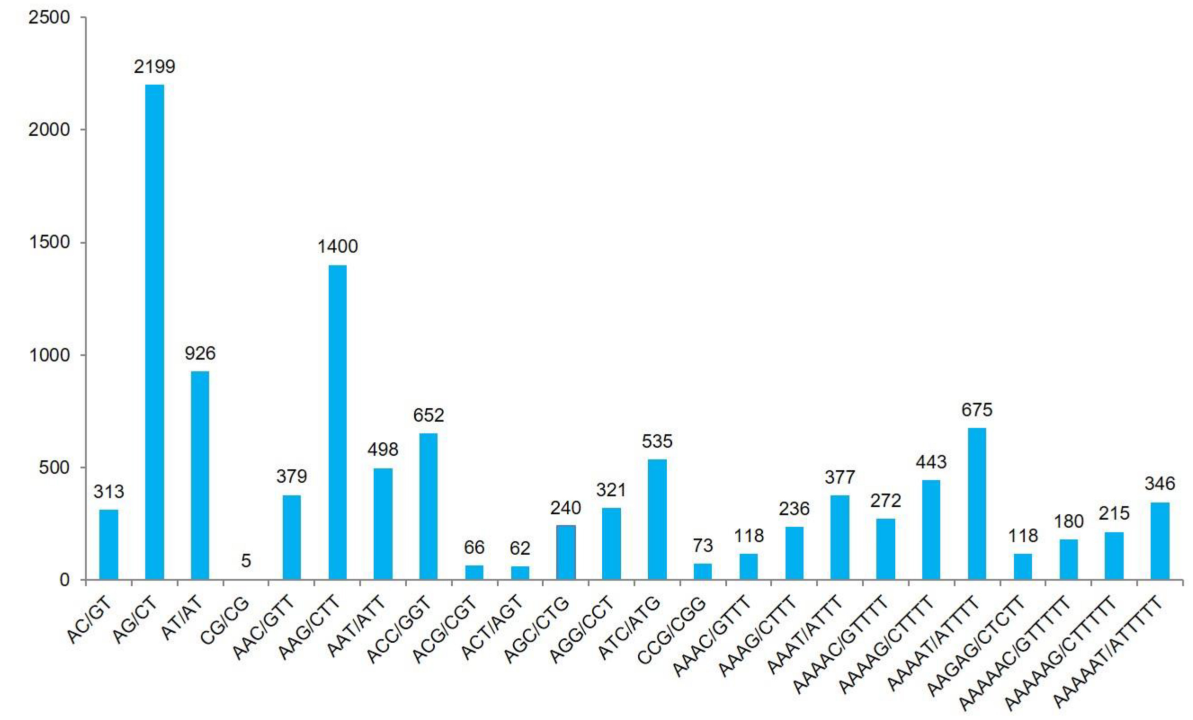


FIGURE 2 | The most abundant types of repeats belonging to dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs and the numbers of SSR motifs.

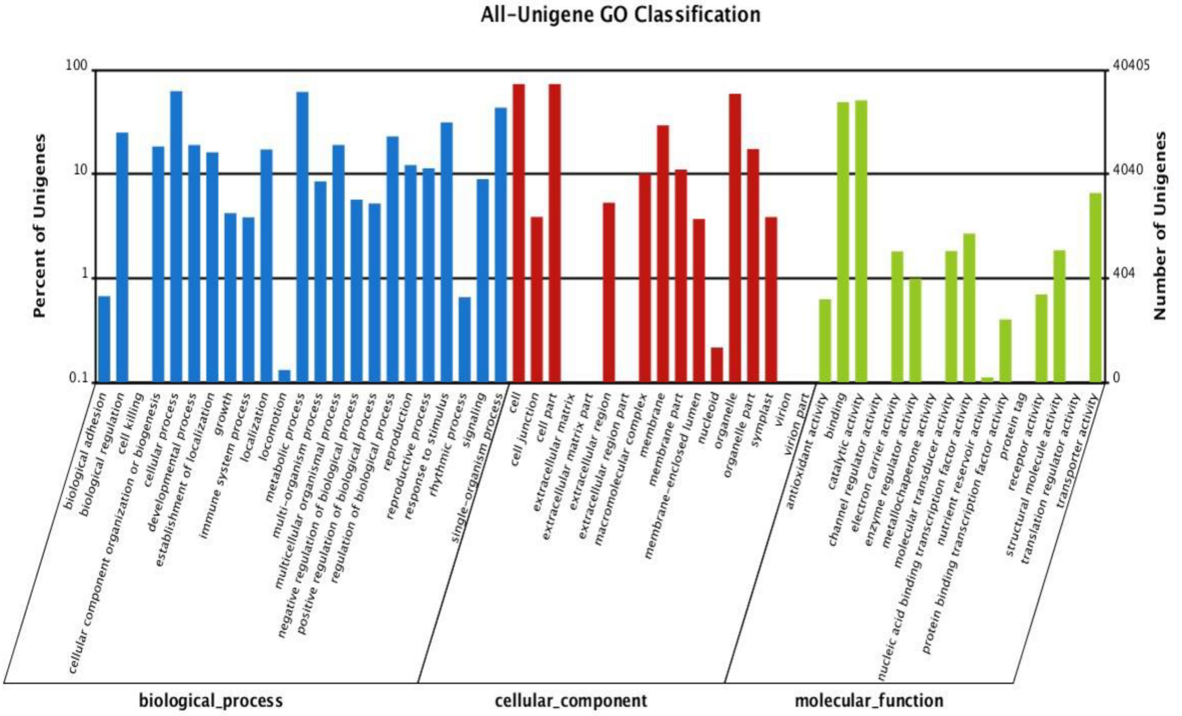


FIGURE 3 | Gene ontology classification of assembled unigenes. The 40,405 matched unigenes were classified into three functional categories: biological process, cellular component, and molecular function.

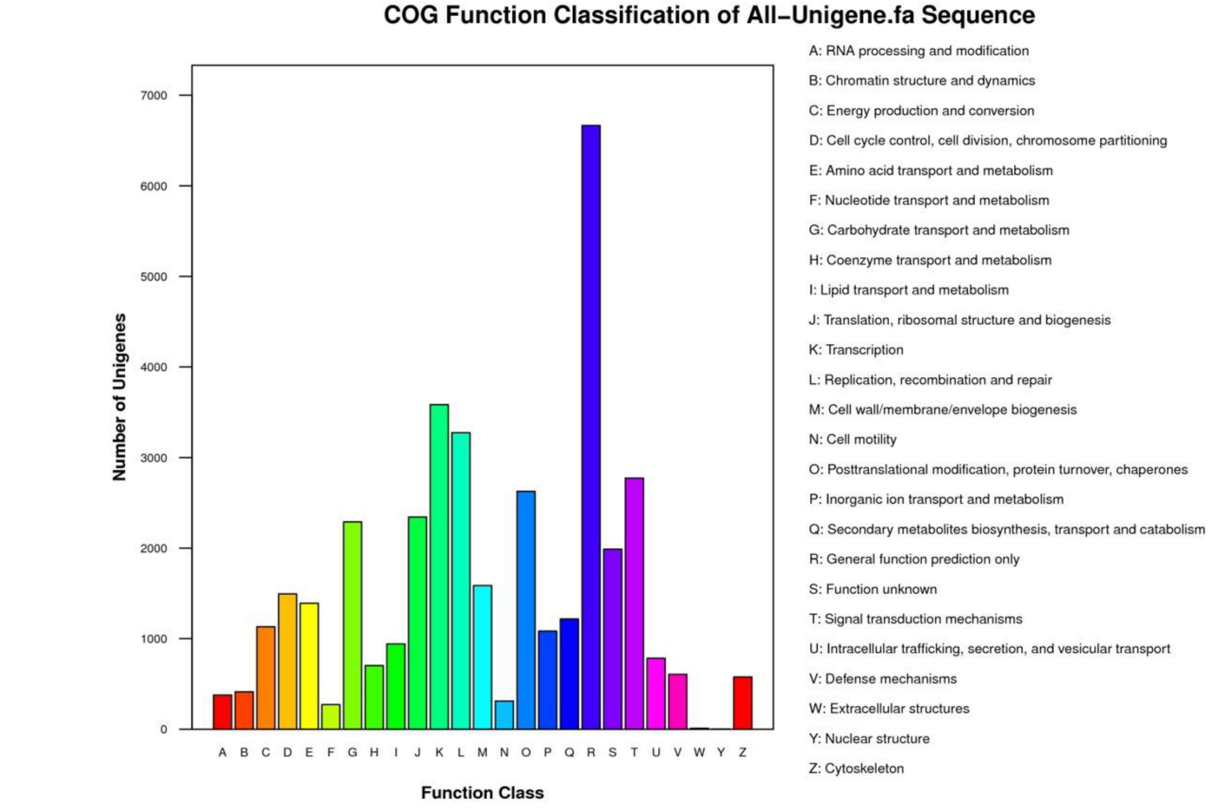


FIGURE 4 | COG classification of all unigenes according to 25 specific categories.

pathways, 30,746 unigenes were categorized into five KEGG biochemical pathways: Cellular Processes (A), Environmental Information Processing (B), Genetic Information Processing (C), Metabolism (D), and Organismal Systems (E). The pathways involving the highest number of unique transcripts were "metabolic pathways" (6,857), followed by "biosynthesis of secondary metabolites" (3,498), "plant-pathogen interaction" (2,329), and "plant hormone signal transduction" (1,738) (**Supplementary Table 11** and **Figure 5**).

Genetic Diversity in *Pistacia*

Diversity analyses were performed in 11 pistachio cultivars (*P. vera*) and in 78 wild *Pistacia* accessions. Allele ranges, number of alleles (Na), effective number of alleles (Ne), PICs, and expected and observed heterozygosities of 55 eSSR loci are given in **Table 3**.

A total of 434 alleles were amplified by 55 eSSR loci, ranging from 4 to 20 per locus. The highest number of allele (Na = 20) was amplified by the CUPVEST3927 locus. The effective number of allele ranged from 1.36 (CUPVEST9032) to 12.31 (CUPVEST3927). The highest observed heterozygosity (Ho = 0.92) was obtained from the CUPVEST1146 locus. The average expected heterozygosity was calculated as 0.65, ranging from 0.26 (CUPVEST9032) to 0.92 (CUPVEST3927). The PICs ranged from 0.25 (CUPVEST9032) to 0.92 (CUPVEST3927), with an average of 0.61 (**Table 4**).

In *P. vera*, a total of 150 alleles were amplified by 55 eSSR loci. The polymorphism rate was 92.7%. A total of 51 eSSR loci were polymorphic, whereas four loci were monomorphic. The average and the highest numbers of alleles were calculated as 2.73 and 6.00 (CUPVEST1313 and CUPVEST4033), respectively. The highest effective number of the allele (Ne = 4.94) was obtained from the CUPVEST4033 locus. The averages of expected heterozygosity, observed heterozygosity, and PIC values were calculated as 0.40, 0.38, and 0.34, respectively (**Table 3** and **Supplementary Table 12**).

In *P. khinjuk*, a total of 152 alleles were produced by 52 polymorphic and 3 monomorphic eSSR loci with 94.6% polymorphism rate. An average of 2.76 alleles were obtained from 52 polymorphic loci, and the highest number of alleles (Na = 5) was in the CUPVEST3826 and CUPVEST4033 loci. The effective number of alleles ranged from 1.00 to 3.85, with an average of 2.17. The highest He (0.74) value was calculated in the CUPVEST1313 and CUPVEST3826 loci. The averages of Ho, He, and PIC values were 0.61, 0.49, and 0.42, respectively (**Table 3** and **Supplementary Table 13**).

The lowest polymorphism (61.8%) was in *P. eurycarpa* with 34 polymorphic SSR loci. The number of alleles ranged from 1 to 4, with a total of 105 alleles. The highest values for Na, Ne, and He in *P. eurycarpa* were obtained from the CUPVEST1313 locus. The expected heterozygosity and PIC values were calculated as 0.33 and 0.27, respectively (**Table 3** and **Supplementary Table 14**).

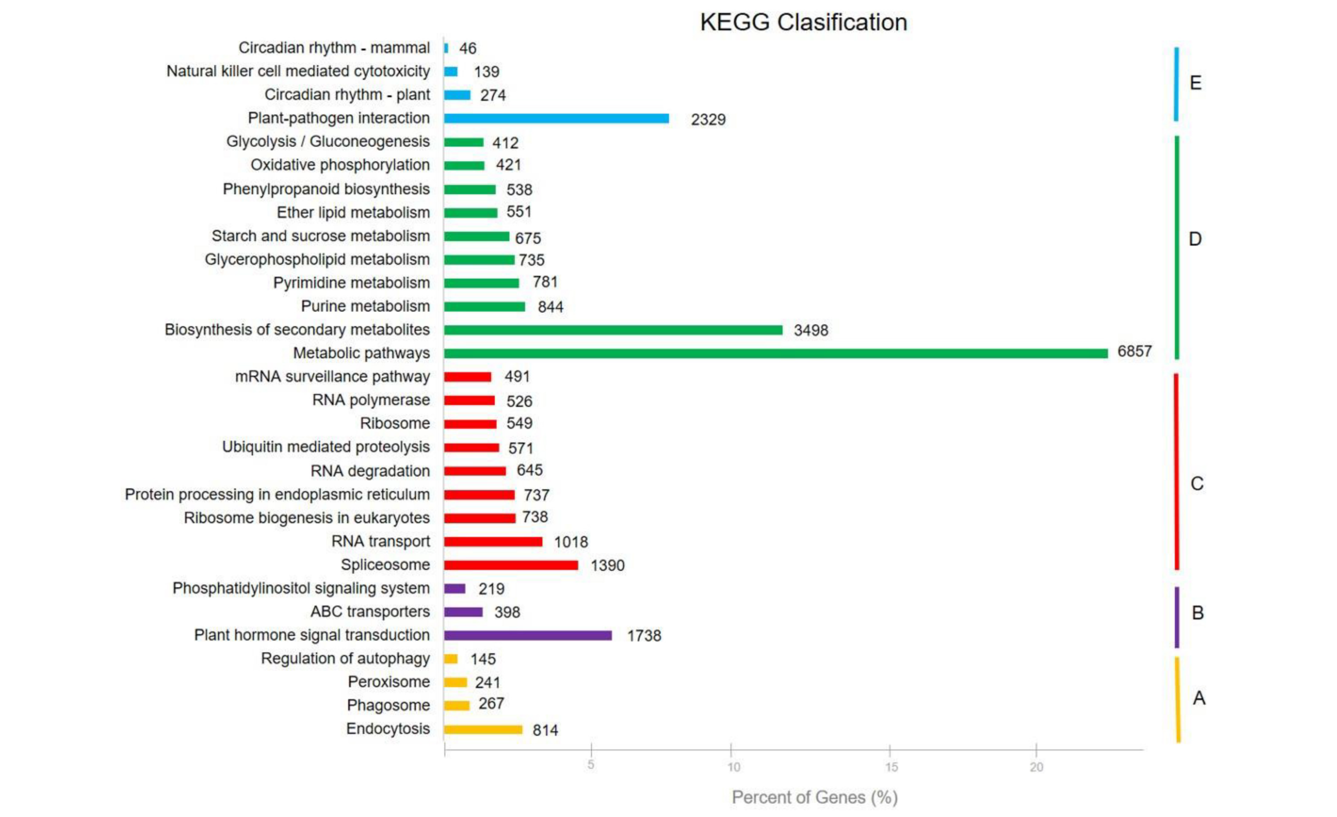


FIGURE 5 | Five KEGG biochemical pathways that obtained from 128 KEGG pathways: **(A)** Cellular Processes, **(B)** Environmental Information Processing, **(C)** Genetic Information Processing, **(D)** Metabolism, and **(E)** Organismal Systems.

TABLE 3 | Mean of population genetic parameters SSR loci in each of *Pistacia* species.

Population	No. of alleles	Polymorphic allele (%)	Polymorphic/monomorphic markers	Na	Ne	Ho	He	PIC
Pistacia	434	100.00%	55/1	7.89	3.40	0.39	0.65	0.61
P. vera	150	92.73%	51/4	2.73	1.93	0.38	0.40	0.34
P. khinjuk	152	94.55%	52/3	2.76	2.17	0.61	0.49	0.42
P. eurycarpa	105	61.82%	34/21	1.91	1.76	0.48	0.33	0.27
P. atlantica	191	94.55%	52/3	3.47	2.12	0.34	0.40	0.36
P. mutica	117	65.45%	36/19	2.13	1.71	0.30	0.30	0.25
UCB-1	109	70.91%	39/16	1.98	1.83	0.56	0.36	0.30
P. integerrima	104	69.09%	38/17	1.89	1.72	0.63	0.34	0.27
P. chinensis	151	90.91%	50/5	2.75	1.82	0.34	0.35	0.31
P. terebinthus	180	83.64%	46/9	3.27	2.02	0.34	0.38	0.33
P. palaestina	114	61.81%	34/21	2.07	1.63	0.28	0.28	0.31
P. lentiscus	131	69.09%	38/17	2.38	1.65	0.24	0.29	0.27
Unknown	163	96.36%	53/2	2.96	2.10	0.45	0.46	0.41

Pistacia atlantica had the highest number of alleles among *Pistacia* species in this study. A total of 191 alleles were obtained from 52 polymorphic and three monomorphic SSR loci. The CUPVEST3927 locus amplified the highest number of allele and the effective number of the allele. The averages of Na and Ne values were 3.47 and 2.12, respectively. The highest Ho value (Ho = 0.92) in *P. atlantica* was calculated in CUPVEST8600 locus, whereas the highest expected heterozygosity (He = 0.88) was calculated in the CUPVEST3927 locus. The expected heterozygosity value ranged from 0.00 to 0.88, with an average of 0.40. The average of PIC value was calculated as 0.36 (**Table 3** and **Supplementary Table 15**).

In *P. mutica*, 117 alleles were obtained from 36 polymorphic SSR loci and 19 monomorphic SSR loci in five accessions. The mean Na and Ne values were determined as 2.13 and 1.71, respectively. The highest observed and expected heterozygosities

TABLE 4 | Novel genic SSR genetic diversity values in 89 *Pistacia* individuals: allele ranges, number of allele (Na), number of effective allele (Ne), observed heterozygosity (Ho), expected heterozygosity (He), and (PIC) values of 55 loci.

SSR loci	Repeat motifs	Allele ranges (bp)	Na	Ne	Ho	He	PIC
CUPVEST1146	(AAAGA)$_4$	150–173	9	3.37	0.92	0.7	0.65
CUPVEST1313	(GGTGGT)$_4$	167–216	12	7.04	0.56	0.86	0.84
CUPVEST1566	(GTT)$_8$	168–190	7	4.80	0.26	0.79	0.76
CUPVEST1855	(TTTAT)$_4$	124–156	9	5.01	0.59	0.80	0.77
CUPVEST1887	(CAT)$_8$	154–179	10	4.68	0.49	0.79	0.75
CUPVEST1981	(ATGGGC)$_4$	105–145	10	3.26	0.50	0.69	0.66
CUPVEST2110	(TCTTC)$_4$	162–205	7	2.48	0.21	0.6	0.53
CUPVEST2125	(ATC)$_8$	151–170	15	5.32	0.57	0.81	0.79
CUPVEST2611	(TGA)$_7$	105–122	6	1.42	0.06	0.30	0.28
CUPVEST2680	(CCTTC)$_4$	93–146	4	1.89	0.28	0.47	0.43
CUPVEST2831	(AAG)$_7$	154–171	9	4.27	0.30	0.77	0.74
CUPVEST3015	(TTCAA)$_4$	127–158	9	3.49	0.26	0.71	0.67
CUPVEST3655	(AAG)$_7$	164–181	8	2.83	0.38	0.65	0.60
CUPVEST3826	(CGG)$_5$	83–136	14	3.97	0.35	0.75	0.72
CUPVEST3917	(AGAAG)$_4$	129–145	6	2.51	0.23	0.60	0.52
CUPVEST3927	(TAG)$_7$	147–179	20	12.31	0.61	0.92	0.92
CUPVEST3929	(TGAG)$_5$	123–136	9	2.54	0.45	0.61	0.56
CUPVEST4033	(TTGT)$_5$	164–179	11	4.01	0.40	0.75	0.71
CUPVEST4068	(GAT)$_8$	138–167	10	3.79	0.55	0.74	0.70
CUPVEST4100	(GTTTA)$_4$	140–241	11	3.73	0.50	0.73	0.67
CUPVEST4140	(CTCA)$_5$	153–168	6	1.75	0.17	0.43	0.41
CUPVEST4279	(GGGGA)$_4$	165–179	7	3.82	0.23	0.74	0.70
CUPVEST4680	(ATCATA)$_4$	165–183	4	2.15	0.27	0.53	0.43
CUPVEST4746	(ATG)$_7$	142–160	8	3.67	0.54	0.73	0.68
CUPVEST5225	(CAT)$_8$	126–165	6	1.48	0.15	0.32	0.32
CUPVEST5301a	(TGGGG)$_4$	161–171	5	3.38	0.43	0.70	0.66
CUPVEST534	(TTGTT)$_4$	116–127	6	2.35	0.50	0.57	0.54
CUPVEST5726	(ATCAC)$_5$	149–175	9	4.96	0.38	0.80	0.77
CUPVEST5746	(GAAGG)$_4$	99–123	5	3.46	0.23	0.71	0.66
CUPVEST5786	(CTC)$_7$	146–162	6	3.64	0.37	0.73	0.68
CUPVEST5836	(TCA)$_8$	124–145	7	2.17	0.19	0.54	0.50
CUPVEST5852	(AACCCT)$_4$	128–147	6	1.53	0.16	0.34	0.31
CUPVEST6009	(CTTTTT)$_4$	112–135	12	4.13	0.48	0.76	0.74
CUPVEST6106	(AGA)$_7$	131–146	5	3.28	0.34	0.70	0.65
CUPVEST6113	(GAA)$_7$	109–124	5	2.67	0.85	0.63	0.55
CUPVEST6536	(GAAGAT)$_4$	136–161	6	2.49	0.36	0.60	0.52
CUPVEST6656	(AAAT)$_6$	115–123	6	3.38	0.42	0.70	0.68
CUPVEST6662	(TTG)$_8$	181–187	4	2.34	0.26	0.57	0.48
CUPVEST6733	(GAA)$_7$	103–179	4	2.37	0.34	0.58	0.48
CUPVEST6938	(ACG)$_7$	172–179	5	2.93	0.48	0.66	0.59
CUPVEST7025	(CT)$_{10}$	157–171	7	4.57	0.49	0.78	0.74
CUPVEST7130	(GTGAGT)$_4$	126–145	5	2.82	0.59	0.65	0.60
CUPVEST7232	(GTGGA)$_4$	151–162	6	3.13	0.41	0.68	0.63
CUPVEST8057	(CAA)$_7$	144–177	12	4.75	0.40	0.79	0.77
CUPVEST8360	(TAG)$_7$	119–134	6	4.15	0.36	0.76	0.73
CUPVEST8592	(AAGGGA)$_4$	166–174	6	2.09	0.33	0.52	0.47
CUPVEST8600	(TGATT)$_5$	160–178	12	2.61	0.63	0.62	0.60
CUPVEST8824	(TTTCT)$_4$	146–167	10	4.78	0.41	0.79	0.76
CUPVEST8845	(GATAAG)$_4$	146–171	13	2.21	0.43	0.55	0.52
CUPVEST901	(TTTCTT)$_4$	134–150	7	3.35	0.27	0.70	0.65
CUPVEST9032	(CTCA)$_5$	96–116	6	1.36	0.25	0.26	0.25

(Continued)

TABLE 4 | Continued

SSR loci	Repeat motifs	Allele ranges (bp)	Na	Ne	Ho	He	PIC
CUPVEST9042	$(AGT)_7$	136–165	7	3.03	0.44	0.67	0.62
CUPVEST921	$(CAGAGC)_4$	140–165	5	2.77	0.40	0.64	0.6
CUPVEST9273	$(AGAAGG)_4$	126–144	7	2.17	0.23	0.54	0.49
CUPVEST9343	$(TTC)_7$	124–138	7	2.85	0.21	0.65	0.58
Total			434				
Mean			7.89	3.40	0.39	0.65	0.61

were calculated as 0.30. The average PIC value was 0.25 in *P. mutica* (**Table 3** and **Supplementary Table 16**).

In *P. atlantica* × *P. integerrima* (UCB1) hybrids, 39 of the 55 eSSR loci were polymorphic, with 70.9% polymorphism. A total of 109 alleles were produced by 55 eSSR loci. The highest number of the allele was obtained from the CUPVEST1146, CUPVEST4746, and CUPVEST8824 loci. The average number of allele and the effective number of allele values were 1.98 and 1.83, respectively. The highest He value was generated by CUPVEST1146, CUPVEST4746, and CUPVEST8824 loci. The average Ho and He values were calculated as 0.56 and 0.36, respectively. The average PIC value in UCB1 seedings was 0.30 (**Table 3** and **Supplementary Table 17**).

In *P. integerrima*, the number of alleles ranged from 1 to 4. A total of 104 alleles were produced by 55 eSSR loci with 69.1% polymorphism. The average number of alleles was calculated as 1.89. The CUPVEST6536 locus amplified the highest effective number of the allele with 3.20. The highest value for He (0.69) was obtained from the CUPVEST6536 locus. The average Ho and He values were 0.63 and 0.34, respectively. The average PIC value was calculated as 0.27 in *P. integerrima* (**Table 3** and **Supplementary Table 18**).

In *P. chinensis*, a total of 151 alleles were generated by 50 polymorphic and 5 monomorphic loci, ranging from 1 to 6 with an average 2.75 alleles per locus. The average Ne, Ho, He, and PIC were calculated as 1.82, 0.34, 0.35, and 0.31, respectively (**Table 3** and **Supplementary Table 19**).

In *P. terebinthus*, 180 alleles were obtained from 46 polymorphic and 9 monomorphic loci. The average number of the allele was 3.27. The highest number of the alleles was obtained from the CUPVEST3927 and CUPVEST6009 loci. The average effective number of allele and the highest effective number of alleles were 2.02 and 6.54, respectively. The CUPVEST3927 SSR locus produced the highest expected heterozygosity value. The average Ho, He, and PIC values were calculated as 0.34, 0.38, and 0.33, respectively (**Table 3** and **Supplementary Table 20**).

In *P. palaestina*, 197 alleles were generated from 34 polymorphic and 21 monomorphic loci. The average number of the allele was detected as 2.07. The average effective number of allele was 1.63. The average Ho, He, and PIC values were calculated as 0.28, 0.28, and 0.31, respectively (**Table 3** and **Supplementary Table 21**).

In *P. lentiscus*, 38 of 55 eSSR loci were polymorphic with 69.1%. A total of 131 alleles were amplified with an average of 2.38 alleles per locus. The average effective number of the allele was detected as 1.65. The highest number of allele (Na = 7), the effective number of the allele (Ne = 3.90), observed heterozygosity (Ho = 0.92), and expected heterozygosity (He = 0.74) values were obtained from the CUPVEST3826, CUPVEST5726, CUPVEST1146, and CUPVEST5726 loci, respectively. The average values for Ho and He were 0.24 and 0.29, respectively. The average PIC value was calculated as 0.27 (**Table 3** and **Supplementary Table 22**).

Genetic diversity values of unknown accessions were calculated as well. A total of 53 eSSR loci were polymorphic with 96.4%. The average numbers of Na, Ne, Ho, He, and PIC values were 2.96, 2.10, 0.45, 0.46, and 0.41, respectively (**Table 3** and **Supplementary Table 23**).

Clustering and Structure Analysis

The maximum Delta *K* value was at *K* = 7 (**Figure 6**). *Pistacia* accessions were grouped in seven main clusters: Cluster 1 included *P. vera* cultivars, whereas *P. khinjuk* was the closest species to *P. vera*. *P mutica* and *P. atlantica* were in Cluster 3 together with *P. eurycarpa*, which was clearly separated from *P. mutica* and *P. atlantica*. Cluster 4 included *P. integerrima* and UCB1 (*P. atlantica* × *P. integerrima*) accessions. Three UCB1 accessions were clearly separated from *P. integerrima*. *P. chinensis* accessions were in Cluster 5. *P. terebinthus* and *P. palaestina* accessions were grouped in the same cluster (Cluster 6) together with unknown *Pistacia* accessions. Cluster 7 consisted *P. lentiscus* accessions that were separated from rest of the accessions in the UPGMA analysis (**Figure 7**). A PCoA supported the structure and UPGMA analysis (**Figure 8**).

DISCUSSION

Identification of eSSRs

The density of eSSRs distribution was computed as one SSR per 17.75 kb in the present study and was higher than other species such as *Argyranthemum brousonetii* (2.3%, 27.00 kb) and *Zingiber officinale* (2.7%, 25.20 kb) (White et al., 2016; Awasthi et al., 2017), while it was lower than in some species such as *Arachis hypogaea* (17.7%, 3.30 kb), *Curcuma longa* (20.5%, 4.80 kb), and *Curcuma alismatifolia* (12.5%, 6.60 kb), respectively (Annadurai et al., 2013; Wang et al., 2018; Taheri et al., 2019). eSSR frequencies and their density in genomes may differ from species to species, because each species has different genetic construction. On the other hand, using different bioinformatics tools and criteria for detection of SSRs may also be a reason for the differences (Liu et al., 2018; Taheri et al., 2019).

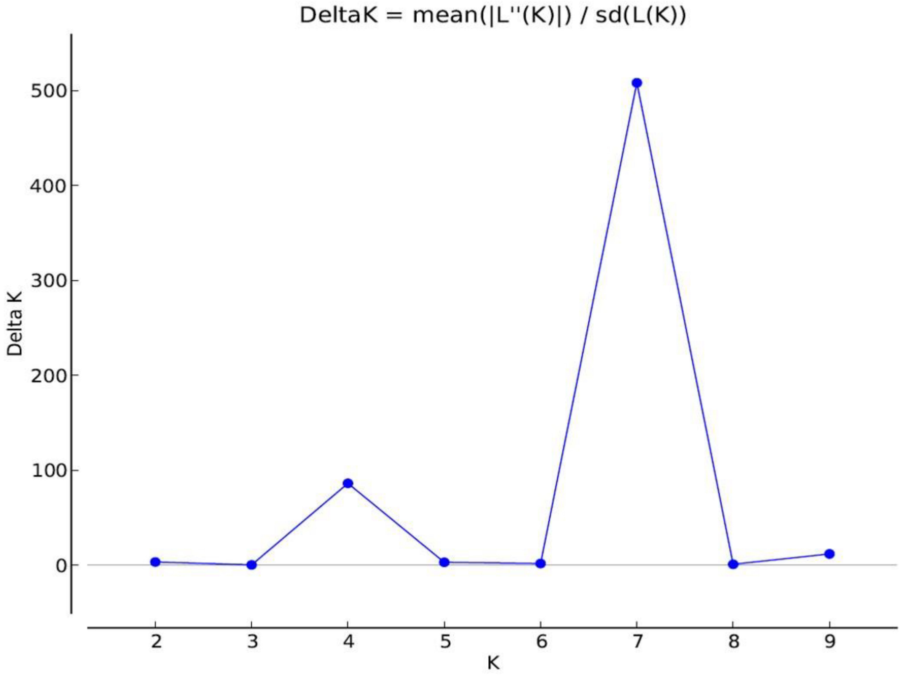


FIGURE 6 | The graph demonstrates number of possible population ($K = 7$) and subpopulation ($K = 4$). X axis refers to number of K, and Y axis refers to Delta K.

eSSR repeat motifs identified from dinucleotide to hexanucleotide and trinucleotide repeats (29.54%), dinucleotide repeats (24.06%), and hexanucleotide repeats (20.67%) were the most abundant repeat motifs, respectively. These results were similar to those found in previous studies in abundance of trinucleotide motifs within the transcriptome sequences (Awasthi et al., 2017; Park et al., 2019; Taheri et al., 2019). A previous study in *P. vera* by Jazi et al. (2017) also demonstrated that dinucleotides (44.7%) and trinucleotides (40.6%) were the most abundant types of repeats. The most frequent types in genic SSRs are trinucleotide repeat type, whereas the common types of SSRs in unigene sequences are dinucleotide and trinucleotide types (Varshney et al., 2002).

Transcriptome Assembly and Functional Annotation

De novo sequencing and assembly without aligning with the reference genome have been widely used to obtain first sequences for non-model organisms (Russell et al., 2014; Wei et al., 2014; Chen et al., 2017). The transcriptome sequences of *P. vera* were the first provided by Jazi et al. (2017) for discovery of markers about salinity-related genes. In the present study, different tissues of a female and a male *P. vera* cultivars were sequenced, and N50 of unigenes was computed as 1,333 bp, similar with the study performed by Jazi et al. (2017).

The GO database is one of the largest information sources about detection of the genes functions ranging from model organisms to minor organisms (Gene Ontology Consortium, 2004). For GO analysis, a total of 40,405 sequences were associated with 325,220 GO terms, which classified different 56 subcategories in three major categories in this study. Jazi et al. (2017) demonstrated that 68,539 sequences were annotated with 302,375 GO terms. The results indicated that assembled unigenes have different molecular functions involved in different metabolic pathways. The assembled sequences were aligned to COG database for prediction of the possible functions. KO provides recognition of the biological pathway of transcripts using KEGG database (Blanca et al., 2011; Torre et al., 2014). Therefore, KEGG pathway analysis explains the information about biological systems of organisms and the relationships between transcripts and their molecular, cellular, and organism levels (Kanehisa et al., 2008). A total of 30,746 unigenes were associated with 279 KEGG pathways. The results illustrated that the KEGG pathway classification of the *P. vera* will facilitate to understand related complex traits at transcriptome level in pistachio and in closely related species.

eSSR Polymorphism in *Pistacia*

SSR markers have been widely preferred for genetic diversity studies, construction of consensus genetic linkage maps, QTL mapping, and MAS in breeding programs (Li et al., 2012; Dong et al., 2018). SSR markers from transcriptome sequences are especially valuable because of their preserved gene regions (Taheri et al., 2019). NGS provides more opportunities than detecting classic SSRs and generates enormous data for development of SSRs in many plant species including *P. vera* (Jazi et al., 2017). The detection of potential SSR markers in pistachio is very easy owing to its high heterozygosity rate (Motalebipour et al., 2016; Jazi et al., 2017).

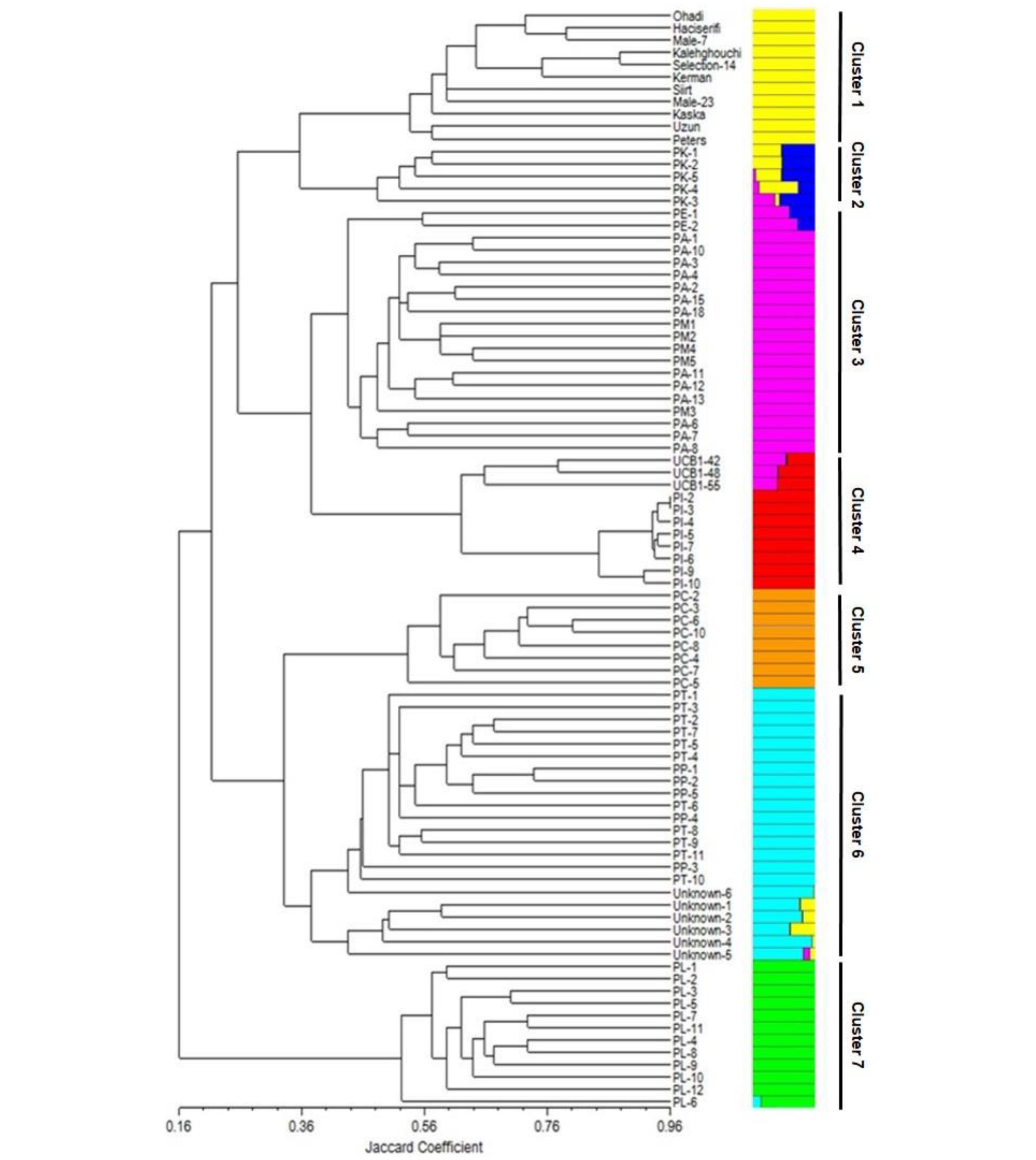


FIGURE 7 | UPGMA dendrogram of 11 *P. vera* and 78 wild *Pistacia* genotypes belongs to *P. khinjuk*, *P. eurycarpa*, *P. atlantica*, *P. mutica*, UCB1, *P. integerrima*, *P. chinensis*, *P. terebinthus*, *P. palaestina*, and *P. lentiscus*. The figure refers to the number of possible populations at *K* = 7 (Delta *K* = 7).

Although Jazi et al. (2017) detected 11,337 potential SSRs in *P. vera*, a total of 14,308 genic SSR loci were determined in this study.

In *P. vera*, several reports were published regarding the development of novel genomic SSR markers in pistachio. First, novel 14 SSR markers were developed by Ahmad et al. (2003)

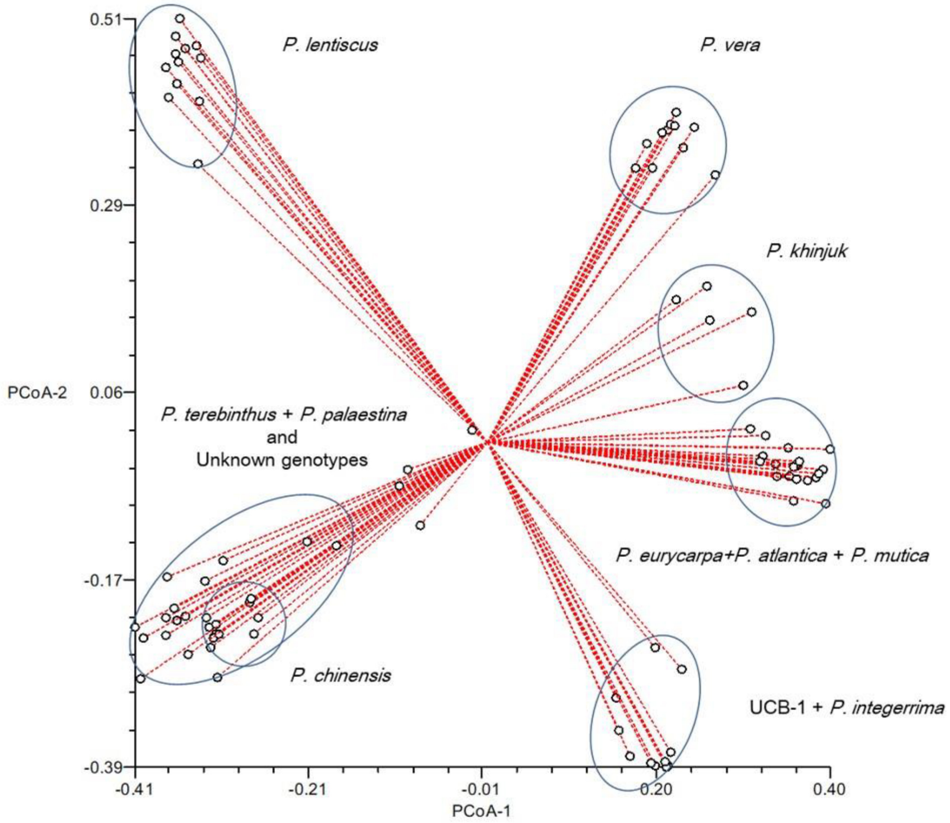


FIGURE 8 | Principal coordinate analysis of the 89 *Pistacia* individuals.

in *P. vera*. Kolahi-Zonoozi et al. (2014) designed 42 primers from Akbari pistachio cultivar and selected 12 polymorphic SSR loci in 45 economically important pistachio cultivars. Then, Zaloğlu et al. (2015) constructed genomic libraries enriched with dinucleotides and trinucleotides repeats. They developed 47 polymorphic SSR loci from *P. vera* cv. Siirt. Topçu et al. (2016) sequenced 192 clones from enriched $(GA)_n$ and $(AAG)_n$ motifs. In total, 110 of 135 primers were produced in PCRs, and 64 of them were found polymorphic with 264 alleles in 12 pistachio cultivars. Currently, there are a few studies about the development of genomic SSR markers in *Pistacia* using NGS technology (Motalebipour et al., 2016; Khodaeiaminjan et al., 2018). Motalebipour et al. (2016) developed SSR markers from *P. vera*, and their transferability was tested in five wild *Pistacia* species. Khodaeiaminjan et al. (2018) developed 625 polymorphic *in silico* SSR markers. We developed here genic SSR markers using NGS and generated 33,341 SSR loci.

Motalebipour et al. (2016) developed 204 SSR markers and tested them in five wild *Pistacia* species. The means of Ho, He, and PIC values in 24 *P. vera* accessions were calculated as 0.46, 0.55, and 0.50, respectively (Motalebipour et al., 2016). Khodaeiaminjan et al. (2018) used 613 *in silico* SSR loci with an average number of allele 4.2 in 18 *P. vera* genotypes. Average values for observed (Ho) and expected (He) heterozygosities and PIC were 0.53, 0.56, and 0.51, respectively, in 18 *P. vera* cultivars. In this study, we used 55 eSSR loci with an average number of 2.73 allele in only *P. vera*.

There are a limited number of studies about development of SSR markers in wild *Pistacia* species. Albaladejo et al. (2008) developed novel SSR markers from *P. lentiscus* using enriched library method and produced 59 alleles from eight SSR markers, ranging from 3 to 13 per locus, although a total of 131 alleles were generated by 55 eSSR loci in *P. lentiscus* genotypes in this study. Arabnezhad et al. (2011) designed 27 primer pairs enriched for dinucleotide $(AG)_n$ and trinucleotide $(ATG)_n$ motifs from *P. khinjuk* sequences. A total of 114 alleles were generated with an average number of allele Na = 2.8 per locus in all *Pistacia* accessions (Arabnezhad et al., 2011). In this study, 52 of 55 eSSRs were found polymorphic, and a total of 152 alleles were generated with an average number of allele Na = 2.8 in five *P. khinjuk* accessions.

Phylogeny of *Pistacia* Species

In this study, structure analysis separated *Pistacia* accessions in seven main clusters. *P. khinjuk* was the closest species to cultivated *P. vera* as in previous studies (Zohary, 1952; Kafkas and Perl-Treves, 2001, 2002; Golan-Goldhirsh et al., 2004; Al-Saghir and Porter, 2006; Kafkas, 2006a; Motalebipour et al., 2016). *P. lentiscus* was the most distant species to *P. vera* in accordance with previous studies (Zohary, 1952; Kafkas and Perl-Treves, 2001, 2002; Golan-Goldhirsh et al., 2004; Kafkas, 2006a,b; Motalebipour et al., 2016).

Zohary (1952) defined that *P. eurycarpa* is a subspecies of *P. atlantica* (var. *kurdica*), whereas it was classified as a different species by Yaltırık (1967). In this study, we used two old *P. eurycarpa* accessions sampled from Göbek village of Gaziantep province to emerge relationship with other *Pistacia* species. The results clearly demonstrated that *P. eurycarpa* was closer to *P. atlantica* than *P. khinjuk*. Kafkas and Perl-Treves (2001) found that *P. atlantica* and *P. eurycarpa* were closely related species at molecular level. Therefore, the results in this study supported statements at morphological level by Yaltırık (1967) and at molecular level by Kafkas and Perl-Treves (2001).

Al Yafi (1978) detected different subspecies of *P. atlantica* using some leaf characters. He reported that *P. mutica* had been a subspecies in *P. atlantica*. Kafkas (2006a) described that *P. atlantica* and *P. mutica* were not clearly diverged. Therefore, *P. mutica* was grouped within *P. atlantica*. In this study, pairs of species were not prominently separated from each other. The previous findings supported that *P. atlantica* and *P. mutica* species were one of the closest pairs of species (Al-Saghir and Porter, 2006; Kafkas, 2006a).

P. integerrima and its hybrids have been used as rootstock for *P. vera* in California. Zohary (1952) defined *P. integerrima* as a subspecies of *P. chinensis*, whereas Parfitt and Badenes (1997) classified it as a distinct species. In this study, *P. integerrima* and *P. chinensis* were clearly separated from each other. Similar results were also obtained by Kafkas and Perl-Treves (2001), Kafkas (2006a), and Motalebipour et al. (2016).

There is still discussion about whether *P. terebinthus* and *P. palaestina* are same or different species. The first classification was performed by Engler (1883), who considered *P. palaestina* as a variety of *P. terebinthus*. However, Zohary (1952) described that *P. palaestina* was a different species from *P. terebinthus*. On the other hand, Yaltırık (1967) reported that *P. palaestina* was a subspecies of *P. terebinthus.* Kafkas and Perl-Treves (2001) and Kafkas (2006a) supported Yaltirik and Engler's classification studies. In this study, these species were not prominently diverged from each other. Therefore, our hypothesis is that they are same species, and *P. palaestina* can be a subspecies of *P. terebinthus*.

In this study, six unknown *Pistacia* genotypes were grouped with *P. palaestina* and *P. terebinthus* according to UPGMA analysis. Structure analysis demonstrated that they have close relationships with cultivated *P. vera*. This situation can be described that these genotypes may be hybrids between *P. vera* and *P. palaestina* or *P. terebinthus*.

CONCLUSION

Transcriptome sequencing provided opportunities for mining easy and cost-effective SSR markers using NGS platform. A total of 98,831 unigenes in this study can be useful for genome annotation in *P. vera* and in related species in the future. The SSR distribution frequency in pistachio transcriptome was one SSR per 17.75 kb. A total of 14,308 eSSRs were defined using transcriptome data of pistachio, and they can be used in studies such as germplasm characterization, population and evolutionary studies, marker-assisted breeding, and association and QTL mapping in *Pistacia* species. This was the first study characterizing 10 *Pistacia* species by genic SSRs and provided important findings on the taxonomy of *Pistacia* species.

DATA AVAILABILITY STATEMENT

The sequence data have been deposited in the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA648340.

AUTHOR CONTRIBUTIONS

HK, EI, and SK prepared plant materials. HK, AP, HT, and SK performed the bioinformatic and SSR analysis. HK, AP, and SK wrote the manuscript. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Çukurova University Scientific Research Projects Unit (Project No: FYL-2016-7326 and FBA-2015-4538).

ACKNOWLEDGMENTS

The authors wish to thank the Pistachio Research Institute and Çukurova University for providing plant materials.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.01021/full#supplementary-material

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M. A. K., Polymeropoulos, M. H., Xiao, H., et al. (1991). Complementary Dna sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656. doi: 10.1126/science.2047873

Ahmad, R., Ferguson, L., and Southwick, S. M. (2003). Identification of pistachio (*Pistacia vera* L.) nuts with microsatellite markers. *J. Am. Soc. Hortic. Sci.* 128, 898–903. doi: 10.21273/jashs.128.6.0898

Al Yafi, J. (1978). New characters differentiating *Pistacia atlantica* subspecies. *Candollea*. 33, 201–208.

Albaladejo, R. G., Sebastiani, F., Aparicio, A., Buonamici, A., González-Martínez, S. C., and Vendramin, G. G. (2008). Development and characterization of eight polymorphic microsatellite loci from *Pistacia lentiscus* L. (*Anacardiaceae*). *Mol. Ecol. Resour.* 8, 904–906. doi: 10.1111/j.1755-0998.2008.02110.x

Al-Saghir, M. G., and Porter, D. M. (2006). Random amplified polymorphic DNA (RAPD) study of *Pistacia* species (*Anacardiaceae*). *Asian J. Plant Sci.* 5, 1002–1006. doi: 10.3923/ajps.2006.1002.1006

Annadurai, R. S., Neethiraj, R., Jayakumar, V., Damodaran, A. C., Rao, S. N., Katta, M. A. V. S. K., et al. (2013). De novo transcriptome assembly (NGS) of *Curcuma longa* L. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. *PLoS One* 8:e56217. doi: 10.1371/journal.pone.0056217

Arabnezhad, H., Bahar, M., and Pour, A. T. (2011). Evaluation of genetic relationships among Iranian pistachios using microsatellite markers developed from *Pistacia khinjuk* Stocks. *Sci. Hortic.* 128, 249–254. doi: 10.1016/j.scienta.2011.01.028

Awasthi, P., Singh, A., Sheikh, G., Mahajan, V., Gupta, A. P., Gupta, S., et al. (2017). Mining and characterization of EST-SSR markers for *Zingiber officinale* Roscoe with transferability to other species of *Zingiberaceae*. *Physiol. Mol. Biol. Plants* 23, 925–931. doi: 10.1007/s12298-017-0472-5

Blanca, J., Cañizares, J., Roig, C., Ziarsolo, P., Nuez, F., and Picó, B. (2011). Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (*Cucurbitaceae*). *BMC Genomics* 12:104. doi: 10.1186/1471-2164-12-104

Chen, C., Zhou, P., Choi, Y. A., Huang, S., and Gmitter, F. G. (2006). Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.* 112, 1248–1257. doi: 10.1007/s00122-006-0226-1

Chen, H., Liu, L., Wang, L., Wang, S., Somta, P., and Cheng, X. (2015). Development and validation of EST-SSR markers from the transcriptome of adzuki bean (*Vigna angularis*). *PLoS One* 10:e0131939. doi: 10.1371/journal.pone.0131939

Chen, J., Li, R., Xia, Y., Bai, G., Guo, P., Wang, Z., et al. (2017). Development of EST-SSR markers in flowering Chinese cabbage (*Brassica campestris* L. ssp. Chinensis var. utilis Tsen et Lee) based on de novo transcriptomic assemblies. *PLoS One* 12:e0184736. doi: 10.1371/journal.pone.0184736

Dong, M., Wang, Z., He, Q., Zhao, J., Fan, Z., and Zhang, J. (2018). Development of EST-SSR markers in Larix principis-rupprechtii Mayr and evaluation of their polymorphism and cross-species amplification. *Trees - Struct. Funct.* 32, 1559–1571. doi: 10.1007/s00468-018-1733-9

Doyle, J. J., and Doyle, J. L. (1987). Genomic plant DNA preparation from fresh tissue-CTAB. *Phytochem. Bull.* 19, 11–15. doi: 10.2307/4119796

Ellis, J. R., and Burke, J. M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity* 99, 125–132. doi: 10.1038/sj.hdy.6801001

Engler, A. (1883). "Burseraceae et Anacardiceae," in *Monopgraphiae Phaneragamarum*, ed. A. C. De- Candolle (Port Jervis, NY: Lubrecht and Cramer Ltd.), 284–293.

FAO (2019). *FAOSTAT, Crops*. Rome: Food and Agriculture Organization of the United Nations.

Ferrão, L. F. V., Caixeta, E. T., Pena, G., Zambolim, E. M., Cruz, C. D., Zambolim, L., et al. (2015). New EST–SSR markers of *Coffea arabica*: transferability and application to studies of molecular characterization and genetic mapping. *Mol. Breed.* 35:31. doi: 10.1007/s11032-015-0247-z

Gadaleta, A., Giancaspro, A., Giove, S. L., Zacheo, S., Mangini, G., Simeone, R., et al. (2009). Genetic and physical mapping of new EST-derived SSRs on the A and B genome chromosomes of wheat. *Theor. Appl. Genet.* 118, 1015–1025. doi: 10.1007/s00122-008-0958-1

Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261. doi: 10.1093/nar/gkh036

Golan-Goldhirsh, A., Barazani, O., Wang, Z. S., Khadka, D. K., Saunders, J. A., Kostiukovsky, V., et al. (2004). Genetic relationships among mediterranean *Pistacia* species evaluated by RAPD and AFLP markers. *Plant Syst. Evol.* 246, 9–18. doi: 10.1007/s00606-004-0132-4

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Gundesli, M. A., Kafkas, S., Zarifikhosroshahi, M., and Kafkas, N. E. (2019). Role of endogenous polyamines in the alternate bearing phenomenon in pistachio. *Turk. J. Agric. For.* 43, 265–274. doi: 10.3906/tar-1807-74

Gupta, P. K., Rustgi, S., Sharma, S., Singh, R., Kumar, N., and Balyan, H. S. (2003). Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics* 270, 315–323. doi: 10.1007/s00438-003-0921-4

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hu, J. B., Zhou, X. Y., and Li, J. W. (2010). Development of novel EST-SSR markers for cucumber (*Cucumis sativus*) and their transferability to related species. *Sci. Hortic.* 125, 534–538. doi: 10.1016/j.scienta.2010.03.021

Huang, H., Lu, J., Ren, Z., Hunter, W., Dowd, S. E., and Dang, P. (2011). Mining and validating grape (*Vitis* L.) ESTs to develop EST-SSR markers for genotyping and mapping. *Mol. Breed.* 28, 241–254. doi: 10.1007/s11032-010-9477-2

Jazi, M., Seyedi, S. M., Ebrahimie, E., Ebrahimi, M., De Moro, G., and Botanga, C. (2017). A genome-wide transcriptome map of pistachio (*Pistacia vera* L.) provides novel insights into salinity-related genes and marker discovery. *BMC Genomics* 18:627. doi: 10.1186/s12864-017-3989-7

Kafkas, S. (2006a). Phylogenetic analysis of the genus *Pistacia* by AFLP markers. *Plant Syst. Evol.* 262, 113–124. doi: 10.1007/s00606-006-0460-7

Kafkas, S. (2006b). "Phylogeny, Evolution and Biodiversity in the Genus *Pistacia* (*Anacardiaceae*)," in *Plant Genome, Biodiversity and Evolution*, Vol. 1, Part C, Phanerogams (Angiosperm-Dicotyledons). eds A. K. Sharma and A. Sharma (Enfield: Science Publishers), 525–557.

Kafkas, S., Kaska, N., Wassimi, A. N., and Padulosi, S. (2006). Molecular characterisation of Afghan pistachio accessions by amplified fragment length polymorphisms (AFLPs). *J. Hortic. Sci. Biotechnol.* 81, 864–868. doi: 10.1080/14620316.2006.11512151

Kafkas, S., and Perl-Treves, R. (2001). Morphological and molecular phylogeny of *Pistacia* species in Turkey. *Theor. Appl. Genet.* 102, 908–915. doi: 10.1007/s001220000526

Kafkas, S., and Perl-Treves, R. (2002). Interspecific relationships in *Pistacia* based on RAPD fingerprinting. *HortScience* 37, 168–171. doi: 10.21273/hortsci.37.1.168

Kafkas, S., Perl-Treves, R., and Kaska, N. (2000). Unusual *Pistacia atlantica* Desf. (*Anacardiaceae*) monoecious sex type in the yunt mountains of the Manisa province of Turkey. *Isr. J. Plant Sci.* 48, 277–280. doi: 10.1092/UFCU-7LF6-T0A3-UXWY

Kafkas, S. (2019). "SSR markers in the genus *Pistacia*," in *Landscape Genetics*, eds M. Minhoto Teixeira Filho, M. C. Fujita, and Rodrigues Nogueira, T. A. (London: IntechOpen).

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Khodaeiaminjan, M., Kafkas, S., Motalebipour, E. Z., and Coban, N. (2018). In silico polymorphic novel SSR marker development and the first SSR-based genetic linkage map in pistachio. *Tree Genet. Genomes* 14:45. doi: 10.1007/s11295-018-1259-8

Kolahi-Zonoozi, S. H., Mardi, M., Zeinalabedini, M., Pirseyedi, S. M., Mahmoodi, P., Tabatabaei, I., et al. (2014). Development of 12 new SSR markers for genetic diversity and structure analysis in pistachio (*Pistacia vera* L.). *J. Hortic. Sci. Biotechnol.* 89, 707–711. doi: 10.1080/14620316.2014.11513141

Li, D., Deng, Z., Qin, B., Liu, X., and Men, Z. (2012). De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics* 13:192. doi: 10.1186/1471-2164-13-192

Liu, J., and Muse, S. V. (2005). PowerMarker: integrated analysis environment for genetic marker data. *Bioinformatics* 21, 2128–2129. doi: 10.1093/bioinformatics/bti282

Liu, S., An, Y., Li, F., Li, S., Liu, L., Zhou, Q., et al. (2018). Genome-wide identification of simple sequence repeats and development of polymorphic SSR markers for genetic studies in tea plant (*Camellia sinensis*). *Mol. Breed.* 38:59. doi: 10.1007/s11032-018-0824-z

Motalebipour, E., Kafkas, S., Khodaeiaminjan, M., Çoban, N., and Gözel, H. (2016). Genome survey of pistachio (*Pistacia vera* L.) by next generation sequencing: development of novel SSR markers and genetic diversity in *Pistacia* species. *BMC Genomics* 17:998. doi: 10.1186/s12864-016-3359-x

Parfitt, D. E., and Badenes, M. L. (1997). Phylogeny of the genus *Pistacia* as determined from analysis of the chloroplast genome. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7987–7992. doi: 10.1073/pnas.94.15.7987

Park, D., Kim, J. H., and Kim, N. S. (2019). De novo transcriptome sequencing and gene expression profiling with/without B-chromosome plants of *Lilium amabile*. *Genomics Inform.* 17:e27. doi: 10.5808/GI.2019.17.3.e27

Peakall, R., and Smouse, P. E. (2012). GenALEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., et al. (2003). TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19, 651–652. doi: 10.1093/bioinformatics/btg034

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., et al. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.* 2, 225–238. doi: 10.1007/BF00564200

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Rohlf, F. J. (2009). *NTSYSpc: Numerical Taxonomy System. Exet. Softw. Version 2.*

Russell, J., Hackett, C., Hedley, P., Liu, H., Milne, L., Bayer, M., et al. (2014). The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. *Mol. Breed.* 33, 835–849. doi: 10.1007/s11032-013-9996-8

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* 18, 233–234. doi: 10.1038/72708

Tabbasam, N., Zafar, Y., and Mehboob, U. R. (2014). Pros and cons of using genomic SSRs and EST-SSRs for resolving phylogeny of the genus Gossypium. *Plant Syst. Evol.* 300, 559–575. doi: 10.1007/s00606-013-0891-x

Taheri, S., Abdullah, T. L., Rafii, M. Y., Harikrishna, J. A., Werbrouck, S. P. O., Teo, C. H., et al. (2019). De novo assembly of transcriptomes, mining, and development of novel EST-SSR markers in *Curcuma alismatifolia* (*Zingiberaceae* family) through Illumina sequencing. *Sci. Rep.* 9:3047. doi: 10.1038/s41598-019-39944-2

Topçu, H., Çoban, N., and Kafkas, S. (2016). Novel microsatellite markers in *Pistacia vera* L. and their transferability across the genus *Pistacia*. *Sci. Hortic.* 198, 91–97. doi: 10.1016/j.scienta.2015.11.012

Torre, S., Tattini, M., Brunetti, C., Fineschi, S., Fini, A., Ferrini, F., et al. (2014). RNA-seq analysis of Quercus pubescensLeaves: de novo transcriptome assembly, annotation and functional markers development. *PLoS One* 9:e112487. doi: 10.1371/journal.pone.0112487

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23, 48–55. doi: 10.1016/j.tibtech.2004.11.005

Varshney, R. K., Thiel, T., Langridge, N. S. P., and Graner, A. (2002). In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell. Mol. Biol. Lett.* 7, 537–546.

Wang, H., Lei, Y., Yan, L., Wan, L., Cai, Y., Yang, Z., et al. (2018). Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences. *Crop J.* 6, 172–180. doi: 10.1016/j.cj.2017.09.007

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., et al. (2010). De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* 11:726. doi: 10.1186/1471-2164-11-726

Wei, L., Li, S., Liu, S., He, A., Wang, D., Wang, J., et al. (2014). Transcriptome analysis of *Houttuynia cordata* Thunb. by illumina paired-end RNA sequencing and SSR marker discovery. *PLoS One* 9:e84105. doi: 10.1371/journal.pone.0084105

White, O. W., Doo, B., Carine, M. A., and Chapman, M. A. (2016). Transcriptome sequencing and simple sequence repeat marker development for three macaronesian endemic plant Species. *Appl. Plant Sci.* 4:1600050. doi: 10.3732/apps.1600050

Yaltırık, F. (1967). *Anacardiaceae*. *Flora Turkey* 2, 544–548.

Yang, S., Zhong, Q., Tian, J., Wang, L., Zhao, M., Li, L., et al. (2018). Characterization and development of EST-SSR markers to study the genetic diversity and populations analysis of Jerusalem artichoke (*Helianthus tuberosus* L.). *Genes Genomics* 40, 1023–1032. doi: 10.1007/s13258-018-0708-y

You, F. M., Huo, N., Gu, Y. Q., Luo, M. C., Ma, Y., Hane, D., et al. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253. doi: 10.1186/1471-2105-9-253

Zalapa, J. E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., et al. (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.*. 99, 193–208. doi: 10.3732/ajb.1100394

Zaloğlu, S., Kafkas, S., Doğan, Y., and Güney, M. (2015). Development and characterization of SSR markers from pistachio (*Pistacia vera* L.) and their transferability to eight *Pistacia* species. *Sci. Hortic.* 189, 94–103. doi: 10.1016/j.scienta.2015.04.006

Zhang, J., Liang, S., Duan, J., Wang, J., Chen, S., Cheng, Z., et al. (2012). De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). *BMC Genomics* 13:90. doi: 10.1186/1471-2164-13-90

Zhang, M., Mao, W., Zhang, G., and Wu, F. (2014). Development and characterization of polymorphic ESTSSR and genomic SSR markers for tibetan annual wild barley. *PLoS One* 9:e94881. doi: 10.1371/journal.pone.0094881

Zhang, Z., Xie, W., Zhao, Y., Zhang, J., Wang, N., Ntakirutimana, F., et al. (2019). EST-SSR marker development based on RNA-sequencing of E. sibiricus and its application for phylogenetic relationships analysis of seventeen *Elymus* species. *BMC Plant Biol.* 19:235. doi: 10.1186/s12870-019-1825-8

Zhou, R., Wu, Z., Jiang, F. L., and Liang, M. (2015). Comparison of gSSR and EST-SSR markers for analyzing genetic variability among tomato cultivars (*Solanum lycopersicum* L.). *Genet. Mol. Res.* 14, 13184–13194. doi: 10.4238/2015.October.26.14

Zhu, Y., Hao, Y., Wang, K., Wu, C., Wang, W., Qi, J., et al. (2009). Analysis of SSRs information and development of SSR markers from walnut ESTs. *J. Fruit Sci.* 26, 394–398.

Zohary, M. (1952). A monographical study of the genus *Pistacia*. *Palest. J. Bot.* 5, 187–228.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

HYPOTHESIS AND THEORY
published: 26 October 2020
doi: 10.3389/fgene.2020.556749

Profiling Alternative 3′ Untranslated Regions in Sorghum using RNA-seq Data

*Min Tu and Yin Li**

Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

Sorghum is an important crop widely used for food, feed, and fuel. Transcriptome-wide studies of 3′ untranslated regions (3′UTR) using regular RNA-seq remain scarce in sorghum, while transcriptomes have been characterized extensively using Illumina short-read sequencing platforms for many sorghum varieties under various conditions or developmental contexts. 3′UTR is a critical regulatory component of genes, controlling the translation, transport, and stability of messenger RNAs. In the present study, we profiled the alternative 3′UTRs at the transcriptome level in three genetically related but phenotypically contrasting lines of sorghum: Rio, BTx406, and R9188. A total of 1,197 transcripts with alternative 3′UTRs were detected using RNA-seq data. Their categorization identified 612 high-confidence alternative 3′UTRs. Importantly, the high-confidence alternative 3′UTR genes significantly overlapped with the genesets that are associated with RNA N^6-methyladenosine (m^6A) modification, suggesting a clear indication between alternative 3′UTR and m^6A methylation in sorghum. Moreover, taking advantage of sorghum genetics, we provided evidence of genotype specificity of alternative 3′UTR usage. In summary, our work exemplifies a transcriptome-wide profiling of alternative 3′UTRs using regular RNA-seq data in non-model crops and gains insights into alternative 3′UTRs and their genotype specificity.

Keywords: crop, sorghum, RNA-seq, 3′ untranslated regions, alternative 3′UTR, transcriptome analysis, mRNA N^6- methyladenosine

OPEN ACCESS

Edited by:
Frederic J.J. Chain,
University of Massachusetts Lowell,
United States

Reviewed by:
Guoli Ji,
Xiamen University, China
Wenbin Mei,
Inari Agriculture, United States

****Correspondence:***
Yin Li
liyin.plant@hotmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *28 April 2020*
Accepted: *30 September 2020*
Published: *26 October 2020*

Citation:
Tu M and Li Y (2020) Profiling Alternative 3′ Untranslated Regions in Sorghum using RNA-seq Data. Front. Genet. 11:556749. doi: 10.3389/fgene.2020.556749

INTRODUCTION

Sorghum bicolor is a crop globally used for food, fodder, and fuel. The importance of sorghum in agriculture and bioenergy is due to its advantages in diversity, genetics, and genomics (Calvino and Messing, 2012; Boyles et al., 2019; Xie and Xu, 2019). Sorghum cultivars are divided into groups based on their usage: grain sorghum, sweet sorghum, forage sorghum, and energy sorghum. Grain sorghum ranks 5th in global cereal production (Boyles et al., 2019); grain sorghum and forage sorghum serve as significant sources for animal feed; sweet sorghum and energy sorghum are promising bioenergy crops for sugar- and lignocellulosic-based biofuels (Mullet et al., 2014; Mathur et al., 2017; Li et al., 2018; Yang et al., 2020). Sorghum has a relatively small diploid genome with several reference assemblies (Paterson et al., 2009; Deschamps et al., 2018; McCormick et al., 2018; Cooper et al., 2019).

RNA-seq has become a ubiquitous tool in shaping our understanding of the functions of the genomic components in plants (Stark et al., 2019). RNA-seq data are not only valuable in hypothesis-driven research but also powerful for data-driven analysis to generate new insights and testable hypotheses, directing future functional studies. Its primary applications are to measure the gene expression levels and to calculate differential gene expression (DGE). Like in model species and major crops, RNA-seq has recently been widely employed in sorghum research and advanced our understanding in many aspects of sorghum including development (Davidson et al., 2012; Kebrom et al., 2017; Turco et al., 2017; Leiboff and Hake, 2019), accumulation of sugar/biomass (McKinley et al., 2016, 2018; Mizuno et al., 2016, 2018; Zhang L. M. et al., 2018 Li et al., 2019a,b; Zhang et al., 2019; Hennet et al., 2020), stress responses and tolerance (Dugas et al., 2011; Gelli et al., 2014; Sui et al., 2015; Fracasso et al., 2016; Yang et al., 2018; Varoquaux et al., 2019), senescence (Johnson et al., 2015; Wu et al., 2016a), and regulation of miRNA and long non-coding RNA (Calvino et al., 2011; Sun et al., 2020).

3′UTR harbors important regulatory elements, *i.e.*, poly(A) signals and *cis*-acting elements. Poly(A) signals determine polyadenylation (PA), while cis-acting elements in 3′UTR interact with *trans*-acting factors, such as miRNAs and RNA-binding proteins (Ji et al., 2014). 3′UTR-mediated gene regulation affects mRNA localization, translation, transport, and stability (Srivastava et al., 2018). The usage of 3′UTR is complex and regulated. Multiple transcripts can be produced from the same gene through alternative splicing (AS), which could lead to alternative 3′UTRs (Abdel-Ghany et al., 2016). Additionally, a transcript can produce several isoforms with different 3′UTRs varied in length through alternative polyadenylation (APA) (Wu et al., 2011). Alternative 3′UTRs and poly(A) sites are dynamically regulated by developmental and environmental conditions and are tissue specific (Fu et al., 2016; Lin et al., 2017; Lorenzo et al., 2017; Sun et al., 2017; Téllez-Robledo et al., 2019).

To characterize the dynamics of 3′UTRs and poly(A) sites at the genome-wide level, extensive efforts have been made both in experimental and bioinformatic approaches. Several methods based on oligo(dT)-priming have been established to sequence mRNA poly(A) tails, including polyA capture (Mangone et al., 2010), sequencing APA sites (Fu et al., 2011), PolyA Site Sequencing (Shepard et al., 2011), and PA-seq (Hafez et al., 2013). Besides these, other specialized sequencing methods to capture the 3′ ends of mRNAs have been developed, such as poly(A)-position profiling by sequencing (Jan et al., 2010) and 3′ region extraction and deep sequencing (Hoque et al., 2013). In particular, PolyA-Tag sequencing has been extensively used for genome-wide APA studies in higher plants, including *Arabidopsis* (Shen et al., 2011; Wu et al., 2011, 2016b; Hong et al., 2018), *Chlamydomonas reinhardtii* (Bell et al., 2016), rice (Fu et al., 2016; Ye et al., 2019; Zhou et al., 2019), and *Medicago* (Wu et al., 2014). In contrast to the experimental approaches targeting poly(A) tails and/or 3′UTRs, many bioinformatic methods and pipelines have been developed to identify poly(A) sites and/or detect differential APA usage between regular RNA-seq samples [reviewed by Chen et al. (2020)]. These bioinformatics tools fall into four major types: type 1 tools include QAPA (Ha et al., 2018) and PAQR (Gruber et al., 2018) and rely on a pre-existed annotation of poly(A) sites, such as PolyA_DB (Lee et al., 2007), Polysite (Gruber et al., 2016), APASdb (You et al., 2014), and PlantAPA (Wu et al., 2016b); type 2 tools depend on RNA-seq-based transcript reconstruction to infer 3′UTRs, including 3USS (Le Pera et al., 2015) and ExUTR (Huang and Teeling, 2017); type 3 tools, such as Kleat (Birol et al., 2015) and ContextMap 2 (Bonfert and Friedel, 2017), identify poly(A) sites using poly(A)-capped reads in RNA-seq data; and type 4 software include a majority of tools, namely, PHMM (Lu and Bushel, 2013), GETUTR (Kim et al., 2015), Change-Point (Wang W. et al., 2014), EBChangePoint (Zhang and Wei, 2016), IsoSCM (Shenker et al., 2015), DaPars (Xia et al., 2014), APAtrap (Ye et al., 2018), and TAPAS (Arefeen et al., 2018). The type 4 methods identify APA dynamics based on modeling read density fluctuation in RNA-seq data. The advantages and the limitations for each type of the bioinformatic tools have been reviewed in detail, and their performance was compared using benchmarking datasets (Chen et al., 2020).

Previous studies focused on the experimental methods for sequencing poly(A) tails have enhanced our understanding of the functions of the core components of poly(A) machinery, cleavage and polyadenylation specificity factor complex (F), including CPSF30, CPSF100, and FIP1. CPSF30 affects poly(A) signal recognition on the near-upstream element and thus the choice of poly(A) sites for many genes in *Arabidopsis* (Thomas et al., 2012; Chakrabarti and Hunt, 2015). CPSF100 affects the poly(A) site choices through the far-upstream element, resulting in transcriptional read-through for many genes (Lin et al., 2017). Another subunit called FIP1 (factor interacting with PAP1), which acts as a bridge between poly(A) polymerase (PAP) and CPSF (Preker et al., 1995), has complex roles in regulating poly(A) site selection and responses to ABA and abiotic stresses in *Arabidopsis* (Téllez-Robledo et al., 2019). While these studies immensely contribute to understanding the dynamics, mechanisms, and variations in plant polyadenylation sites, these methods require specialized library-prep protocols and bioinformatic pipelines. Unlike routine RNA-seq, they are technically challenging and more expensive, preventing from becoming widely available for plant scientists. Many plant scientists work on agriculturally and economically important crops with large and complex genomes [*e.g.*, maize (Jiao et al., 2017), sorghum (Paterson et al., 2009; McCormick et al., 2018), wheat (International Wheat Genome Sequencing Consortium [IWGSC], 2008), and sugarcane (Zhang J. et al., 2018)], which present potential technical difficulties when performing multi-omics experiments. Although sequencing costs are rapidly decreasing, the majority of published RNA-seq data available in plants, especially those in non-model crops, were produced using the Illumina short-read sequencing platforms (Stark et al., 2019). For example, in sorghum, a large amount of RNA-seq studies were conventional short-read sequencing used for DGE analysis, with a few analyzing AS (Abdel-Ghany et al., 2016). These situations suggest that 3′UTR analysis using conventional RNA-seq data should be valuable and significant for non-model plants.

In this study, we sought to gain insights into alternative 3′UTRs in sorghum. Due to the importance of sorghum stem as a carbon reservoir and a conduit for the mobilization of water, nutrients, and signaling molecules, we chose to analyze the RNA-seq data of sorghum stems, which was previously generated by our group in three genetically related but phenotypically contrasting lines—Rio, BTx406, and R9188—to identify genes and regulatory networks associated with stem sugar accumulation in sweet sorghum (Li et al., 2019a,b). Our present study has the following objectives: (1) to characterize alternative 3′UTRs in sorghum stems, (2) to gain insights into 3′UTR regulation, and (3) with the identified alternative 3′UTRs, to generate testable hypotheses for future in-depth functional studies of 3′UTR-mediated regulation in sorghum.

MATERIALS AND METHODS

RNA-seq for Profiling 3′ Untranslated Regions

The RNA-seq dataset (NCBI accession PRJNA413691) used for 3′UTR analysis was generated previously by us to study genes and networks associated with soluble sugar accumulation in internodes (internodes 2, 3, and 4, numbered from top to bottom) at four post-stem elongation stages from three sorghum genotypes, Rio, BTx406, and R9188 (Li et al., 2019a). The three genotypes contrast in the phenotypes of stem sugar accumulation: Rio accumulates high contents of sugar in stems during the post-flowering stages (∼20% Brix in the stem-extracted juice), while BTx406 has a low stem sugar content (<10% Brix), with R9188 having an intermediate stem sugar content especially during the post-flowering stages (Li et al., 2019a). R9188 is an introgression line developed from the BTx406/Rio cross followed by one backcross to sweet sorghum Rio and contains the early flowering and dwarf loci introgressed from grain sorghum BTx406 (Ritter et al., 2004). The dataset has an advantage such that the three genetically related genotypes allow us to investigate inter-genotype 3′UTR variations.

For each genotype, the internode samples were collected at four time points, namely, T1, T2, T3, and T4 (flag leaf stage, 100% flowering, 10 days after flowering, and 15 days after flowering, respectively), in the experimental field under a split plot design. RNA was extracted with the TRIzol method, processed to libraries, and sequenced using 150-bp pair-end according to standard protocols. Regular quality control and filtering steps were applied to the raw reads and then mapped to the sorghum reference genome BTx623 (v2.1) with Tophat (Paterson et al., 2009; Trapnell et al., 2012). To calculate gene expression (in reads per kilobase of exon per million mapped sequence reads, RPKM) and differentially expressed genes (DEGs), uniquely mapped reads were used, and DEGs were determined with DEseq and edgeR (Anders and Huber, 2010; McCarthy et al., 2012). Moreover, 18,275, 19,727, and 19,102 genes were expressed in Rio, BTx406, and R9188, respectively (total number of the expressed genes for all timepoints, genic reads per replicate ≥ 10, average RPKM per sample ≥ 1).

Transcriptome-Wide Analysis of Alternative 3′UTRs

The publicly available bioinformatic method priUTR (*P*rogram for *R*NA-seq-based *I*dentification of Alternative 3′*UTR*) was used in this study to identify alternative 3′UTRs using regular RNA-seq data from sorghum (Tu, 2020). The program and its user manual are available at https://github.com/mint1234/3UTR-. priUTR is a type-2 bioinformatic method for identifying alternative 3′UTRs, which relies on RNA-seq-based transcript reconstruction to infer mRNA 3′ end. Detailed information about this method is provided in **Supplementary Figure S1**. The alternative 3′UTR results for each genotype and timepoint were obtained with priUTR after mapping the sorghum RNA-seq data to the BTx623 reference genome using a reference-guided mode with Tophat. The reference-guided transcriptome assembly (in GTF format) and gene expression data generated from the Tophat-Cufflinks pipeline were used as input files for the priUTR program. Only those alternative 3′UTRs detected in all three replicates per genotype and time point were kept. Hereafter, a gene or a transcript with an alternative 3′UTR is mentioned as an alternative 3′UTR gene or an alternative 3′UTR transcript.

We integrated the BTx406-introgressed genes/alleles to find out alternative 3′UTR genes associated with the genetic relationship among R9188, Rio, and BTx406 (Li et al., 2019a). A total of 1,805 genes introgressed from BTx406 into R9188 were previously identified based on RNA-seq-derived SNPs with SAMTools, followed by a series of stringent filters, including the mapping quality, read depth, bi-allelism, and homogeneity and homozygosity of the SNPs. The RNA-seq read mapping data were examined for the alternative 3′UTR genes and compared between Rio, BTx406, and R9188 using a genome browser to identify the alternative 3′UTRs that were specifically detected in Rio but not in BTx406/R9188 or *vice versa*.

Expression of Sorghum m^{6}A Functional Factors

The functional factors (“reader,” “writer,” and “eraser” proteins) of m^{6}A modifications in *Arabidopsis* were reported previously (Duan et al., 2017; Ruzicka et al., 2017; Arribas-Hernández et al., 2018; Scutenaire et al., 2018; Wei et al., 2018), with their maize homologs phylogenetically identified (Miao et al., 2020). The protein sequences of the m^{6}A functional factors were used for a BLAST search in the sorghum genome, with a threshold of E value $\leq 10^{-5}$. The best BLAST hits in sorghum were also verified by the gene orthology between maize and sorghum (Zhang et al., 2017).

Functional Annotation and Enrichment Analysis

To comprehensively understand the functions of alternative 3′UTR genes, the functional annotation containing three major sources of annotation, Gene Ontology (GO), MapMan, and KEGG, was employed (Li et al., 2019a). The GO annotation was from Phytozome and AgriGO (Du et al., 2010). For MapMan, the second-level bincodes of the annotation was used (Usadel et al., 2009). Enriched functional terms of the high-confidence alternative 3′UTR geneset (group 4, see “RESULTS”) or the m^{6}A associated geneset were calculated using clusterProfiler (*hypergeometric* < 0.05; Yu et al., 2012).

RESULTS

Transcriptome-Wide Identification of Alternative 3′UTRs in Sorghum Using RNA-seq Data

We performed a transcriptome-wide analysis of alternative 3′UTRs using the sorghum internode RNA-seq datasets (Li et al., 2019a). A total of 1,371 transcripts were identified to have alternative 3′UTRs for all the three genotypes. Among them, 754, 776, and 848 transcripts derived from 750, 771, and 841 genes were identified in Rio, BTx406, and R9188, respectively (**Figures 1A,B**). We focus on analyzing the features of the alternative 3′UTR genes in terms of their 3′UTR lengths, gene order, and functions rather than the 3′UTR dynamics over the time points. Therefore, the alternative 3′UTR transcripts from the four time points were pooled together for downstream analysis. The read depth within a mRNA tends to drop rapidly when reaching to its 3′ end in RNA-seq data, which is known as the 3′ bias of RNA-seq (Li et al., 2015). Due to the 3′ bias of RNA-seq, we removed any transcripts of which the predicted alternative 3′UTR differed from its corresponding annotation, with no more than 50 bp in length to avoid potential bioinformatic artifacts. After that, 1,197 alternative 3′UTRs were kept. Because priUTR is a type-2 bioinformatic tool for alternative 3′UTR detection and could have inherent limitations of transcript assembly tools, the read mapping results of these 1,197 alternative 3′UTRs were visualized on Integrative Genomics Viewer (IGV; Thorvaldsdóttir et al., 2013) and manually compared with their corresponding annotations in BTx623 reference genome in order to find out potential sources of false prediction and to identify high-confidence alternative 3′UTR (**Figures 1A,C**). Based on the types of possible mis-prediction, the 1,197 transcripts were categorized into five groups (**Figure 1A**). Group 1 (namely, the "3′ overlapping gene" group) represents the alternative 3′UTRs that could be likely mis-predicted because a gene is located at the 3′ downstream of the alternative 3′UTR gene and overlaps with the predicted 3′UTR (**Figure 2A**). Group 2 (namely, the "3′ adjacent gene" group) represents the alternative 3′UTRs that could be mis-predicted because a gene locates at the 3′ downstream of the alternative 3′UTR gene and overlaps only with the predicted 3′UTR extension but not with the annotated 3′UTR region (**Figure 2B**). Group 3 (namely, the "mixed transcript" group) stands for those potentially false predictions, likely due to gene models producing multiple transcripts. In such a situation, the alternative 3′UTR transcript could be embedded within a longer transcript or be mis-predicted due to other transcripts with a similar 3′UTR length (**Figure 2C**). Group 4 is a collection of high-confidence alternative 3′UTR transcripts of which the results in IGV supported the predicted alternative 3′UTRs (**Figures 2D,E**). Group 5 are the transcripts of which the alternative 3′UTR seems to be mis-predicted by other miscellaneous reasons. Groups 1, 2, 3, 4, and 5 contains 96, 179, 91, 612, and 219 transcripts, respectively (**Figure 1A**). Examples of the genome browser views for groups 1, 2, 3, and 4 are shown in **Figures 2A–D**, respectively.

A comparison of the alternative 3′UTR lengths from three sorghum genotypes with those of the annotation showed significantly longer 3′UTRs in our results (**Figure 2F**). Moreover, a comparison of the alternative 3′UTR lengths between the five groups indicated that groups 1 and 2 have longer alternative 3′UTRs than the other groups, matching with our findings that many of them are likely false prediction due to the inclusion of transcribed neighboring genes (**Figure 2G**). Besides that, the "mixed transcript" group (group 3) has significantly higher numbers of transcripts per gene (based on the genome annotation) when compared with the remaining four groups, in agreement with the possible source of false prediction as "mixed transcripts" (**Figure 2H**). Additionally, a comparison of the expression levels between the five groups showed that the expression levels of the "3′ overlapping" genes (group 1), but not other groups, were lower than those of the high-confidence alternative 3′UTRs (group 4), suggesting that high expression levels may not be a major reason for the high-confidence prediction of alternative 3′UTRs (**Supplementary Figure S3**). This manual verification step highlights that the high-confidence alternative 3′UTRs consist of a large fraction of all predictions (group 4, 612 out of 1,197) and identifies several major sources of false prediction.

Association Between Alternative 3′UTRs and RNA m^{6}A Modification

Since APA and RNA m^{6}A modification are two molecular phenomena that could be associated with changes in 3′UTR in plants, we therefore sought to test whether the alternative 3′UTRs identified here are associated with APA or RNA m^{6}A modification.

APA is one important approach currently known to produce length variations in 3′UTR in plants. Alternative polyadenylation and/or 3′UTRs are regulated by genetic factors, environmental conditions, and developmental context. Polyadenylation (PA) is directly controlled by the CPSF complex that contains six subunits, CPSF30, CPSF73, CPSF100, CPSF160, Wdr33, and FIP1, and recognizes the polyadenylation sites on 3′UTR (Mandel et al., 2008; Shi et al., 2009). Some subunits of the PA machinery (CPSF30, CPSF100, and FIP1) have been functionally characterized in plants (Thomas et al., 2012; Lin et al., 2017; Téllez-Robledo et al., 2019). FPA is another *trans*-acting factor regulating the 3′ end formation of mRNA in *Arabidopsis* (Duc et al., 2013). Besides that, abiotic stresses, such as dehydration and salt, can also induce alternative polyadenylation (Sun et al., 2017; Téllez-Robledo et al., 2019). Recently, a comprehensive profiling of APAs in rice across several developmental stages and tissues marks tissue specificity in APA patterns (Fu et al., 2016). Based on the abovementioned observation, it can be hypothesized that, if the 612 high-confidence alternative 3′UTR transcripts (group 4) use their alternative 3′UTRs are due, at least partly, to the previously studied PA machinery proteins or environmental

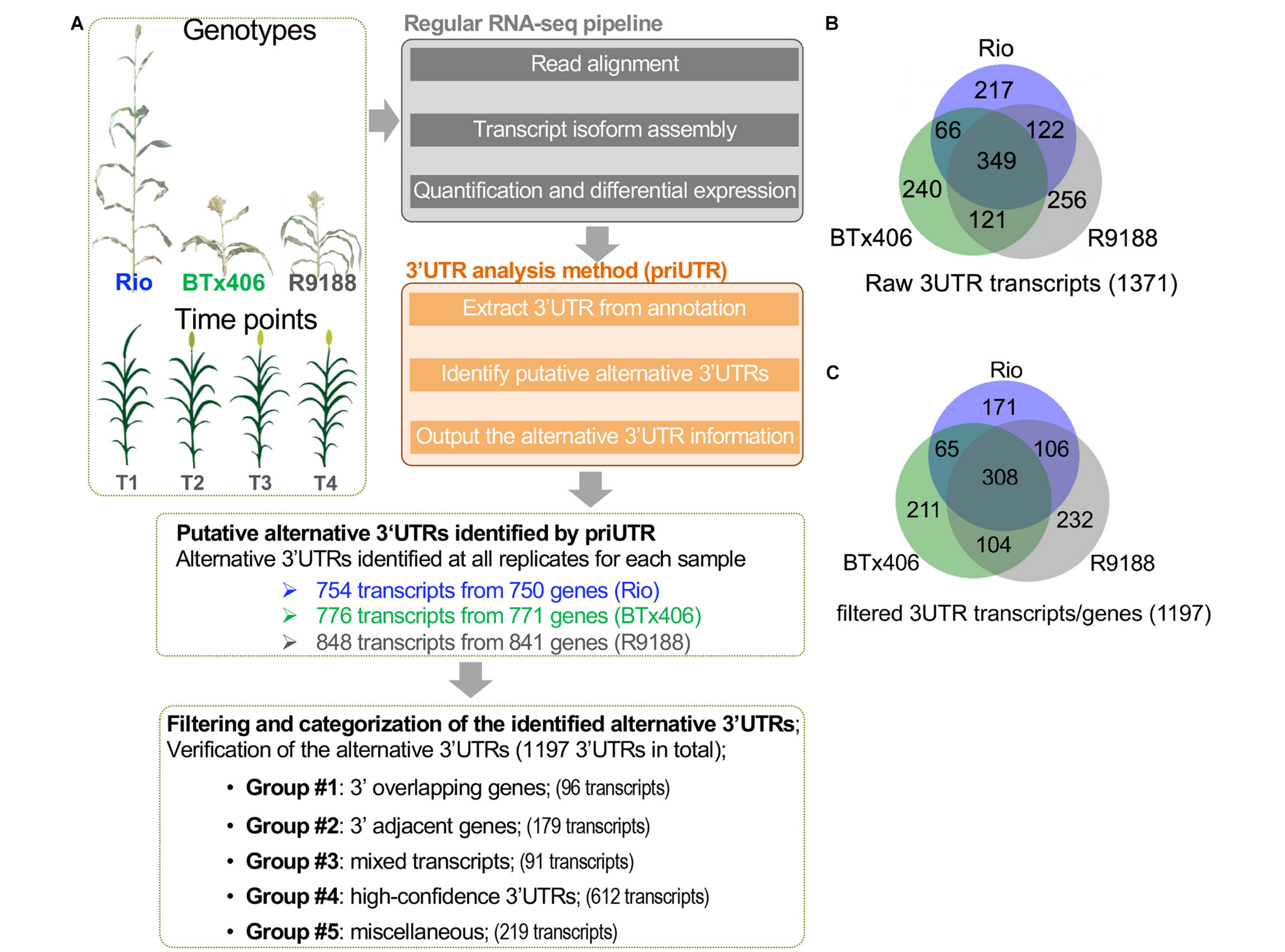

A
Genotypes
Rio
BTx406
R9188
Time points
T1
T2
T3
T4
Regular RNA-seq pipeline
Read alignment
Transcript isoform assembly
Quantification and differential expression
3'UTR analysis method (priUTR)
Extract 3'UTR from annotation
Identify putative alternative 3'UTRs
Output the alternative 3'UTR information
Putative alternative 3'UTRs identified by priUTR
Alternative 3'UTRs identified at all replicates for each sample
754 transcripts from 750 genes (Rio)
776 transcripts from 771 genes (BTx406)
848 transcripts from 841 genes (R9188)
Filtering and categorization of the identified alternative 3'UTRs;
Verification of the alternative 3'UTRs (1197 3'UTRs in total);
Group #1: 3' overlapping genes; (96 transcripts)
Group #2: 3' adjacent genes; (179 transcripts)
Group #3: mixed transcripts; (91 transcripts)
Group #4: high-confidence 3'UTRs; (612 transcripts)
Group #5: miscellaneous; (219 transcripts)
B
Rio
217
66
122
349
240
256
121
BTx406
R9188
Raw 3UTR transcripts (1371)
C
Rio
171
65
106
308
211
232
104
BTx406
R9188
filtered 3UTR transcripts/genes (1197)

FIGURE 1 | Characterization of the alternative 3′ untranslated regions (3′UTRs) in sorghum internodes. **(A)** A diagram showing the sorghum 3′UTR analysis workflow using the published RNA-seq datasets of sorghum internodes at four post-elongation time points. The RNA-seq datasets were from an introgression line R9188 and its two parental lines Rio and BTx406, with contrasting phenotypes in plant height, flowering time, and stem sugar accumulation. 3′UTR analysis was performed using the publicly available 3′UTR tool (priUTR) to identify putative alternative 3′UTRs compared to the sorghum reference genome (cultivar BTx623, v2.1). A total of 1,371 predicted alternative 3′UTRs transcripts were initially identified, with 754, 776, and 848 transcripts detected in Rio, BTx406, and R9188, respectively. Further filtering steps refined to 1,197 transcripts, with at least 50-bp 3′UTR differences in length between the annotation and prediction. The 1,197 alternative 3′UTRs were categorized into five groups, leading to the identification of 612 high-confidence alternative 3′UTRs (designated as group 4). The other groups may possibly be false predicted due to several interfering factors, such as 3′ overlapping genes (designated as group 1), 3′ adjacent genes (designated as group 2), and mixed transcripts from the same gene (designated as group 3) (see details in "MATERIALS AND METHODS" and **Figure 2**). **(B)** Venn diagram of the 1,371 raw alternative 3′UTR transcripts detected in Rio, BTx406, and R9188. **(C)** Venn diagram of the 1,197 alternative 3′UTR transcripts detected in Rio, BTx406, and R9188, with more than 50-bp length difference compared to the corresponding 3′UTRs in the annotation.

conditions, a significant fraction of the 612 sorghum transcripts would be the orthologs of the APA genes in *Arabidopsis* or rice which are associated with PA proteins or stresses. To test the APA hypothesis, we retrieved the APA-associated genesets (genesets A1, A2, B, D1, D2, D3, E, and G) from the above-mentioned studies in *Arabidopsis* and rice. The details of these genesets are provided in **Figure 3A** and **Supplementary Tables S2, S3**. The gene IDs in *Arabidopsis* and rice were converted to sorghum gene IDs using the established gene orthologous relationships between *Arabidopsis* and sorghum (Van Bel et al., 2018) and those between rice and sorghum (Sakai et al., 2013), respectively. Hypergeometric tests failed to reveal any significant enrichment of the APA-associated genesets within the five groups of sorghum alternative 3′UTR genes ($P < 0.05$; **Figure 3A** and **Supplementary Figure S4**), not supporting the APA hypothesis that the alternative 3′UTR genes observed here are largely contributed by the known polyadenylation machinery proteins or the stresses.

We then analyzed whether the alternative 3′UTRs identified in sorghum could be related to RNA m^6A modification. RNA m^6A modification is another factor known to be related to 3′UTR, as a large portion of m^6A sites in plants are located within mRNA 3′UTR regions (Luo et al., 2014, 2020). m^6A is installed by m^6A methyltransferase [known as "m^6A writer,"

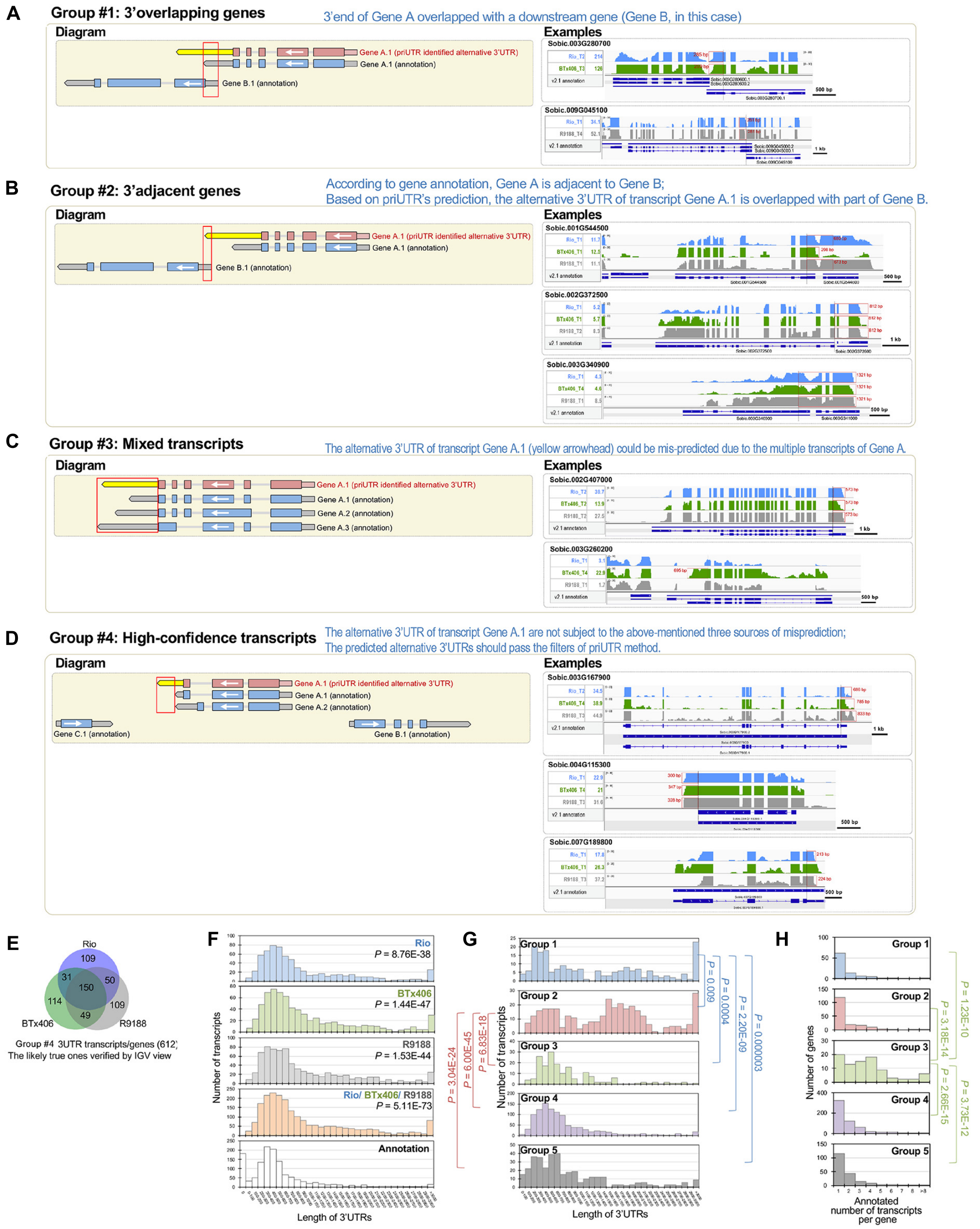
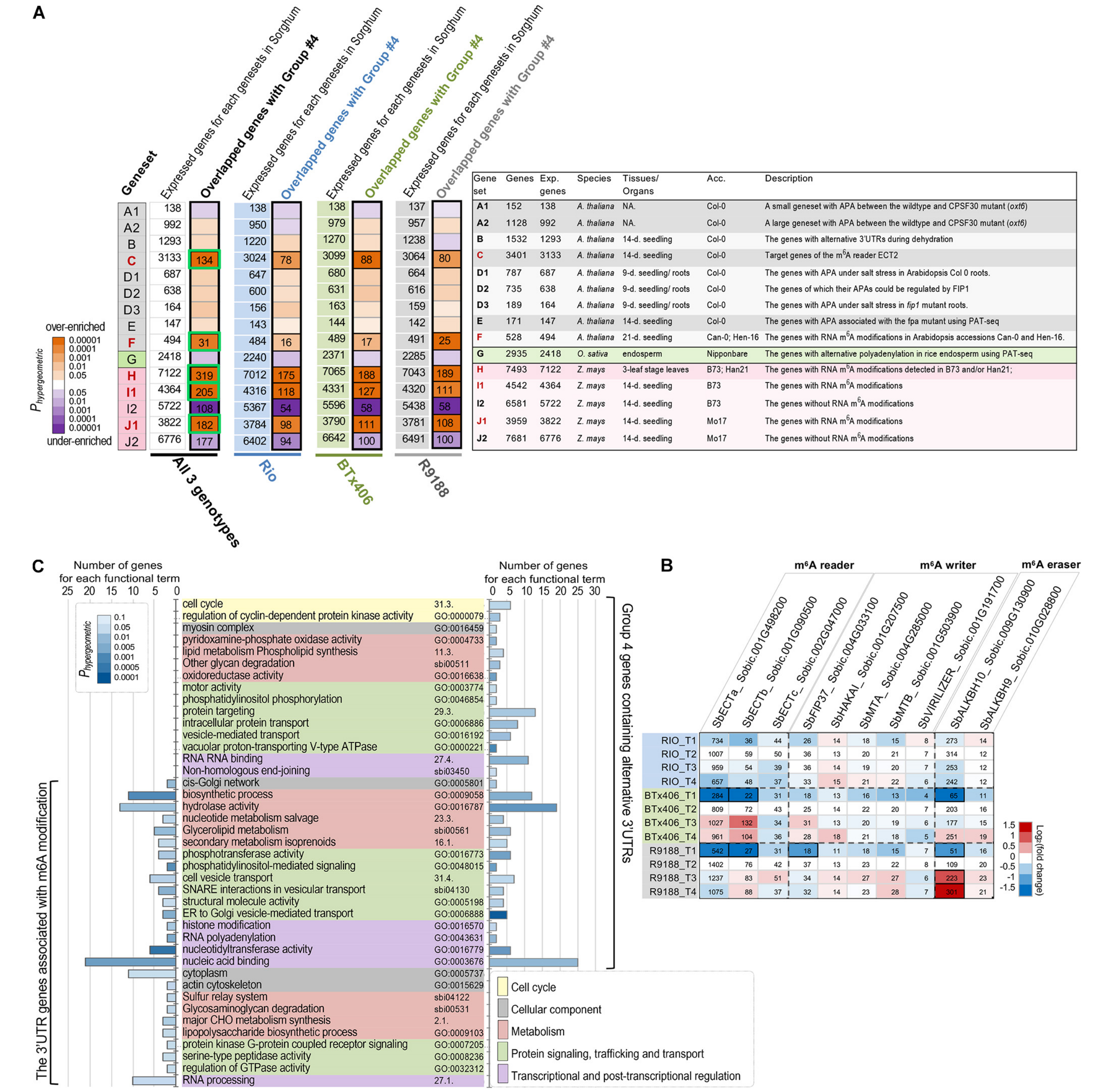


FIGURE 2 | Categorization of the predicted 3′UTRs highlights the high-confidence group of alternative 3′UTRs in sorghum. **(A–D)** Schematic diagrams showing that the predicted 3′UTRs in sorghum were categorized into groups, with groups 1, 2, 3, and 4, respectively. In the left panel of the diagrams, the white arrow indicates gene orientation. The red gene model denotes the reconstructed transcript, while the blue gene model denotes the annotated transcript, with the yellow arrow denoting the identified alternative 3′UTR. In the right panel, representative Integrative Genomics Viewer (IGV) views of the predicted alternative 3′UTRs in groups 1, 2, 3, and 4 (**A–D**, respectively) are shown. The select IGV tracks of the genotype and time point are shown, in which the alternative 3′UTRs were detected. The gene expression levels at those samples provided (in reads per kilobase of exon per million mapped sequence reads). The alternative 3′UTRs are highlighted in red boxes, with their length labeled. **(E)** Venn diagram of the 612 high-confidence alternative 3′UTR transcripts. **(F)** Histograms showing the length of the high-confidence alternative 3′UTRs detected in Rio, BTx406, and R9188, all of the three genotypes and their corresponding annotations. The results indicated that the identified alternative 3′UTRs tend to be longer than the annotations (Wilcoxon rank sum test, $P < 0.05$). **(G)** The histograms of the alternative 3′UTR length of the five groups demonstrated that groups 1 and 2 differed significantly in 3′UTR length from the other groups (Wilcoxon rank sum test, $P < 0.05$), while the length distribution of 3′UTRs was not different among groups 3, 4, and 5. Longer 3′UTRs in groups 1 and 2 are consistent with the classification that the alternative 3′UTRs in groups 1 and 2 could be mostly mis-detected due to the 3′ downstream genes overlapping with or adjacent to the 3′ ends of the alternative 3′UTR genes, respectively. **(H)** Histograms of the number of transcripts per gene for the five groups showed that group 3 has significantly more transcripts per gene than the other groups (Wilcoxon rank sum test, $P < 0.05$), matching the likely source of false positive for group 3 that the alternative 3′UTRs could be mis-predicted by mixing multiple 3′UTRs of the transcripts from the same gene.



Geneset	All 3 genotypes: Expressed genes for each genesets in Sorghum	All 3 genotypes: Overlapped genes with Group #4	Rio: Expressed genes for each genesets in Sorghum	Rio: Overlapped genes with Group #4	BTx406: Expressed genes for each genesets in Sorghum	BTx406: Overlapped genes with Group #4	R9188: Expressed genes for each genesets in Sorghum	R9188: Overlapped genes with Group #4
A1	138		138		138		137	
A2	992		950		979		957	
B	1293		1220		1270		1238	
C	3133	134	3024	78	3099	88	3064	80
D1	687		647		680		664	
D2	638		600		631		616	
D3	164		156		163		159	
E	147		143		144		142	
F	494	31	484	16	489	17	491	25
G	2418		2240		2371		2285	
H	7122	319	7012	175	7065	188	7043	189
I1	4364	205	4316	118	4331	127	4320	111
I2	5722	108	5367	54	5596	58	5438	58
J1	3822	182	3784	98	3790	111	3781	108
J2	6776	177	6402	94	6642	100	6491	100

Gene set	Genes	Exp. genes	Species	Tissues/ Organs	Acc.	Description
A1	152	138	*A. thaliana*	NA.	Col-0	A small geneset with APA between the wildtype and CPSF30 mutant (*oxt6*)
A2	1128	992	*A. thaliana*	NA.	Col-0	A large geneset with APA between the wildtype and CPSF30 mutant (*oxt6*)
B	1532	1293	*A. thaliana*	14-d. seedling	Col-0	The genes with alternative 3'UTRs during dehydration
C	3401	3133	*A. thaliana*	14-d. seedling	Col-0	Target genes of the m^{6}A reader ECT2
D1	787	687	*A. thaliana*	9-d. seedling/ roots	Col-0	The genes with APA under salt stress in Arabidopsis Col 0 roots.
D2	735	638	*A. thaliana*	9-d. seedling/ roots	Col-0	The genes of which their APAs could be regulated by FIP1
D3	189	164	*A. thaliana*	9-d. seedling/ roots	Col-0	The genes with APA under salt stress in *fip1* mutant roots.
E	171	147	*A. thaliana*	14-d. seedling	Col-0	The genes with APA associated with the fpa mutant using PAT-seq
F	528	494	*A. thaliana*	21-d. seedling	Can-0; Hen-16	The genes with RNA m^{6}A modifications in Arabidopsis accessions Can-0 and Hen-16.
G	2935	2418	*O. sativa*	endosperm	Nipponbare	The genes with alternative polyadenylation in rice endosperm using PAT-seq
H	7493	7122	*Z. mays*	3-leaf stage leaves	B73; Han21	The genes with RNA m^{6}A modifications detected in B73 and/or Han21;
I1	4542	4364	*Z. mays*	14-d. seedling	B73	The genes with RNA m^{6}A modifications
I2	6581	5722	*Z. mays*	14-d. seedling	B73	The genes without RNA m^{6}A modifications
J1	3959	3822	*Z. mays*	14-d. seedling	Mo17	The genes with RNA m^{6}A modifications
J2	7681	6776	*Z. mays*	14-d. seedling	Mo17	The genes without RNA m^{6}A modifications

	SbECTa_ Sobic.001G498200	SbECTb_ Sobic.001G090500	SbECTc_ Sobic.002G047000	SbFIP37_ Sobic.004G033100	SbHAKAI_ Sobic.001G207500	SbMTA_ Sobic.004G285000	SbMTB_ Sobic.001G503900	SbVIRILIZER_ Sobic.001G191700	SbALKBH10_ Sobic.009G130900	SbALKBH9_ Sobic.010G028800
RIO_T1	734	36	44	26	14	18	15	8	273	14
RIO_T2	1007	59	50	36	13	20	21	7	314	12
RIO_T3	959	54	39	36	14	19	20	7	253	12
RIO_T4	657	48	37	33	15	21	22	6	242	12
BTx406_T1	284	22	31	18	13	16	13	4	65	11
BTx406_T2	809	72	43	25	14	22	20	7	203	16
BTx406_T3	1027	132	34	31	13	20	19	6	177	15
BTx406_T4	961	104	36	28	18	21	18	5	251	19
R9188_T1	542	27	31	18	11	18	15	7	51	16
R9188_T2	1402	76	42	37	13	23	22	8	109	20
R9188_T3	1237	83	51	34	14	27	27	6	223	23
R9188_T4	1075	88	37	32	14	23	28	7	301	21

cell cycle	31.3.
regulation of cyclin-dependent protein kinase activity	GO:0000079
myosin complex	GO:0016459
pyridoxamine-phosphate oxidase activity	GO:0004733
lipid metabolism Phospholipid synthesis	11.3.
Other glycan degradation	sbi00511
oxidoreductase activity	GO:0016638
motor activity	GO:0003774
phosphatidylinositol phosphorylation	GO:0046854
protein targeting	29.3.
intracellular protein transport	GO:0006886
vesicle-mediated transport	GO:0016192
vacuolar proton-transporting V-type ATPase	GO:0000221
RNA RNA binding	27.4.
Non-homologous end-joining	sbi03450
cis-Golgi network	GO:0005801
biosynthetic process	GO:0009058
hydrolase activity	GO:0016787
nucleotide metabolism salvage	23.3.
Glycerolipid metabolism	sbi00561
secondary metabolism isoprenoids	16.1.
phosphotransferase activity	GO:0016773
phosphatidylinositol-mediated signaling	GO:0048015
cell vesicle transport	31.4.
SNARE interactions in vesicular transport	sbi04130
structural molecule activity	GO:0005198
ER to Golgi vesicle-mediated transport	GO:0006888
histone modification	GO:0016570
RNA polyadenylation	GO:0043631
nucleotidyltransferase activity	GO:0016779
nucleic acid binding	GO:0003676
cytoplasm	GO:0005737
actin cytoskeleton	GO:0015629
Sulfur relay system	sbi04122
Glycosaminoglycan degradation	sbi00531
major CHO metabolism synthesis	2.1.
lipopolysaccharide biosynthetic process	GO:0009103
protein kinase G-protein coupled receptor signaling	GO:0007205
serine-type peptidase activity	GO:0008236
regulation of GTPase activity	GO:0032312
RNA processing	27.1.

FIGURE 3 | The alternative 3′UTR genes in sorghum are associated with m^{6}A modification. **(A)** Significant overlapping between the high-confidence alternative 3′UTR genes in sorghum and the APA- or m^{6}A- associated genesets ($P < 0.01$, hypergeometric test). The significant overlaps between genesets are labeled in black boxes, and the number of overlapping genes is shown. Green box highlights group 4 alternative 3′UTR genes potentially associated with m^{6}A modifications, which is used for functional enrichment analysis **(C)**. The right panel is a summary list of the published genesets of alternative polyadenylation, alternative 3′UTR, or RNA m^{6}A modifications in *Arabidopsis*, rice, and maize. The number of sorghum genes (column "Genes"), expressed genes in sorghum (column "Exp. Genes") orthologous to those in *Arabidopsis*, rice, or maize, and the original tissues/organs and accessions (column "Acc.") are given (details of these genesets are provided in **Supplementary Table S2**). The backgrounds of the genesets are colored according to the original species, and the m^{6}A-associated genesets (C, H, I1, and J1) are highlighted in red. The genesets and their corresponding references are as follow: genesets A1/A2—Thomas et al., 2012; geneset B—Sun et al., 2017; geneset C—Wei et al., 2018; genesets D1/D2/D3—Téllez-Robledo et al., 2019; geneset E - Duc et al., 2013; geneset F—Luo et al., 2014; genesets G—Fu et al., 2016; geneset H—Miao et al., 2020; genesets I1/I2/J1/J2—Luo et al., 2020. NA., not available; 9-days, 9-day old; 14-days, 14-day old; 21-days, 21-day old). **(B)** Expression dynamics of the sorghum genes encoding m^{6}A readers, writers, and erasers. The heatmap is shaded to reflect the magnitude of log2 fold change of gene expression relative to the T2 timepoint per each genotype. Gene expression values in reads per kilobase of exon per million mapped sequence reads are given. The differentially expressed genes per genotype are highlighted in black boxes. **(C)** Functional enrichment results with GO, MAPMAN, and KEGG terms show that both high-confidence alternative 3′UTR genes (group 4) and its subset m^{6}A-associated genes cover a diverse range of function aspects. Only the significant terms ($P < 0.05$, hypergeometric test) with at least two hit genes per term are shown.

e.g., methyltransferase-like 3 (METTL3), METTL14, Wilms tumor 1-associated protein (WTAP); Liu et al., 2014; Wang Y. et al., 2014; Ping et al., 2014], recognized by "m^6A reader" (EVOLUTIONARILY CONSERVED C-TERMINAL REGION proteins, ECT2, 3, 4 in plants; Wei et al., 2018) and removed by "eraser" m^6A RNA demethylase [*e.g.*, alkylated DNA repair protein AlkB homolog 5 (ALKBH5); Zheng et al., 2013]. Previously, mRNA m^6A methylome studies in *Arabidopsis* (Luo et al., 2014), rice (Zhang et al., 2019), and maize (Luo et al., 2020; Miao et al., 2020) reveal a functional association of m^6A with chloroplast, sporogenesis, stress response, and translational status, highlighting new functions of m^6A functional factors, the effects of m^6A modification, and m^6A inter-variety variation. In the present study, if the 612 alternative 3′UTRs in sorghum are associated with or regulated by m^6A modification, a significant fraction of the 612 transcripts would be orthologs of the genes with m^6A modifications in *Arabidopsis* or maize (designated as the m^6A hypothesis). To investigate the hypothesis, several m^6A-associated genesets were obtained from the m^6A-modified genes reported in *Arabidopsis* (genesets C and F) and maize (geneset H, I1, and J1; **Figure 3A**; Luo et al., 2014, 2020; Miao et al., 2020). The genesets I2 and J2 exclusively contain the expressed genes without m6A modification in maize seedlings, serving as the controls for genesets I1 and J1, respectively (Luo et al., 2020). Our analysis clearly showed that the high-confidence alternative 3′UTR genes (group 4) are significantly over-enriched with the orthologs of m^6A-associated genesets in both *Arabidopsis* and maize but are under-enriched with the m^6A control genesets (I2 and J2) ($P_{hypergeometric} < 0.01$; **Figure 3A**). For example, many genes which are orthologous targets of m^6A reader ECT2 in *Arabidopsis* were significantly over-enriched in group 4. The hypergeometric tests were first performed using all of the 612 alternative 3′UTR genes pooled from three sorghum genotypes. We further repeated the analysis in each of the three sorghum lines to avoid potential statistical artifacts. The results for each sorghum genotype confirmed the over-enrichment of m^6A-associated genes (**Figure 3A** and **Supplementary Tables S4, S5**). The statistical significance of the enrichment was further assessed using 400 permutation tests for each of the three sorghum genotypes. In group 4, there are 340, 344, and 358 genes from Rio, BTx406, and R9188, respectively. To avoid potential influences of expression levels in the permutation tests, we examined the distribution of expression levels of the 612 alternative 3′UTR genes within each sorghum genotype (**Supplementary Table S5**). The deciles of the expression levels of group 4 genes were determined in Rio, BTx406, and R9188, respectively, and used to randomly select 36 expressed genes within each expression decile per genotype from all of the expressed genes, yielding a set of 360 random genes with a distribution of expression levels similar to that of the group 4 genes. The results of permutation tests showed that both the number of overlapped genes and the hypergeometric *p*-values were significantly correlated with m^6A genesets C, F, H, I1, and J1 (**Supplementary Figures S5, S6**). In contrast, the number of overlapped genes and the hypergeometric *p*-values were significantly smaller than the random distributions for m^6A control genesets I2 and J2 (**Supplementary Figures S5, S6**), supporting a significant under-enrichment of non-m^6A modified genes in the group 4 genes. Overall, the analysis between genesets is suggestive of the m^6A hypothesis.

To ascertain m^6A occurrence in sorghum internodes, we found that multiple genes putatively encoding m^6A reader, writer, and eraser, respectively, were expressed in the RNA-seq samples. with many of them expressed at high levels. The genes encoding m^6A readers (Sobic.001G498200 and Sobic,001G090500) and eraser (Sobic.009G130900) were upregulated during post-anthesis stages in BTx406 and R9188 (**Figure 3B**).

To portrait the representative molecular functions, biological processes, and metabolic pathways, functional enrichment was performed for the alternative 3′UTR genes and the subset of 3′UTR genes associated with m^6A modifications. The subset of m^6A-associated 3′UTR genes was generated by combining the overlapping genes across several m^6A genesets (green boxes in **Figure 3A**, a total of 398 genes). **Figure 3C** reveals that both the alternative 3′UTR genes and the m^6A-associated 3′UTR genes cover a diverse set of functions, which can be grouped into five major categories, including cell cycle, cellular components, metabolism, protein signaling, trafficking and transport, and transcriptional and post-transcriptional regulation. The enrichment results of these two genesets highlight functions related to metabolism (*i.e.*, biosynthetic process, hydrolase activity) and nucleic acid regulation (*i.e.*, nucleic acid binding and RNA processing). Moreover, we found that the sorghum m^6A-associated 3′UTR genes are highly enriched in RNA processing, nucleic acid binding, as well as functions associated with protein localization and transportation (for example, cell vesicle transport, SNARE in vesicle transport, *cis*-Golgi network) (**Figure 3C**), consistent with the representative functions of m^6A-modified genes previously seen in *Arabidopsis* and maize (Luo et al., 2014, 2020).

Genotype-Specific Alternative 3′UTRs

Utilizing the genetic relationship among Rio, BTx406, and R9188, we sought to address whether alternative 3′UTR can be genotype-specific. The RNA-seq read mapping data of the 612 alternative 3′UTRs were compared between the three genotypes. Twenty-three genes were identified to have distinct alternative 3′UTRs either between Rio and BTx406/R9188 (seven genes, **Supplementary Figure S7A**) or between Rio/R9188 and BTx406 (16 genes, **Supplementary Figures S7B,C**). Many of them (19 out of 23) are homologous to the genes with known annotations or proven important functions in *Arabidopsis*. The genotype-specific alternative 3′UTRs comprise only a small fraction of the identified alternative 3′UTRs (∼3.7%, 23 out of 612). To further validate the parental origins of these 23 alternative 3′UTR genes in R9188, we utilized the 1,805 genes that are introgressed from BTx406 into R9188 (Li et al., 2019a). All of the seven genes in R9188 with identical 3′UTR length to BTx406 are indeed introgressed from BTx406, while among the 16 R9188 genes with identical 3′UTR length to Rio, eight are validated to be originated from Rio, with the remaining eight genes lacking SNPs for validation (**Supplementary Figure S7**). The representative genes in R9188 that have genotype-specific alternative 3′UTRs are shown in **Figure 4**. Many homologs of the genes identified here are critical for plant development.

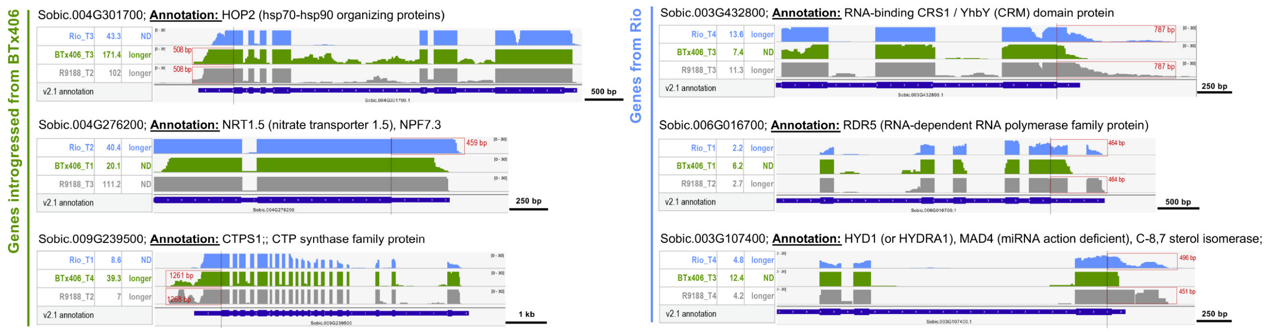


FIGURE 4 | The alternative 3′UTRs in sorghum can be genotype specific. Representative examples of genotype-specific alternative 3′UTRs. Based on the SNPs identified using RNA-seq reads from Rio, BTx406, and R9188, the genes in R9188 were identified to be originated from Rio or introgressed from BTx406. Examples of alternative 3′UTRs distinct between Rio and BTx406/R9188 (the R9188 alleles also introgressed from BTx406): a longer 3′UTR of Sobic.004G276200 was detected only in Rio but not in BTx406/R9188, while longer 3′UTRs of Sobic.004G301700 and Sobic.009G239500 were seen only in BTx406/R9188 but not in Rio. Examples of alternative 3′UTRs distinct between Rio/R9188 and BTx406 (the R9188 alleles originated from Rio): longer 3′UTRs of Sobic.003G432800, Sobic.006G016700, and Sobic.003G107400 were detected in Rio/R9188 but not in BTx406. The select Integrative Genomics Viewer tracks of the genotype and time point are shown, in which the alternative 3′UTRs were detected. The gene expression levels of those samples are provided (in reads per kilobase of exon per million mapped sequence reads). The alternative 3′UTRs are highlighted in red boxes, with their length labeled. ND, not detected.

For example, Sobic.004G276200 is a homolog of AtNRT1.5 (AT1G32450) that regulates root architecture and leaf senescence through nitrate response and root-to-shoot nitrate transport and potassium translocation (Meng et al., 2015; Zheng et al., 2016). Sobic.006G016700 is homologous to AtHYD1 (AT1G20050), a sterol biosynthetic gene that affects cell wall synthesis, phytohormone signaling, and miRNA activity, crucial for embryo and root development (Schrick et al., 2004; Souter et al., 2004; Brodersen et al., 2012).

DISCUSSION

Our work exemplifies a transcriptome-wide 3′UTR analysis using conventional RNA-seq data in non-model plants, identifying several hundred genes with high-confidence alternative 3′UTRs in sorghum. While many bioinformatic tools for identifying mRNA 3′ ends have been reported, the priUTR program was used here to analyze sorghum 3′UTRs for several reasons (Tu, 2020). First, the genome assembly and annotation files of non-model crop species, such as maize and sorghum, can be readily used as input files of priUTR (*e.g.*, gff3 annotation), while some other programs appear to lack clear guidelines for preparing input files using non-model plant species. Second, results from the priUTR tool should be considered largely reliable, since priUTR resembles a previous online tool for detecting alternative 3′UTRs, 3USS, in the principles of the algorithms (Le Pera et al., 2015). priUTR adopts the idea of intron-chain matching from 3USS, which required the transcript reconstruction from Cufflinks or Scripture as one of the input files and predicted alternative 3′UTRs once all of the intron chain could be matched between the assembled and annotated transcript. However, 3USS provided alternative 3′UTR prediction for model animals and *Arabidopsis* and is no longer operational. Unlike 3USS, priUTR utilizes partial intron-chain matching algorithm, in which three consecutive introns before the stop codon are required to be matched and holds the flexibility that the assembled transcript could have alternative splicing in its 5′ part or have additional unannotated introns in the predicted 3′UTR regions (**Supplementary Figure S1**). Third, the priUTR tool could be used for analyzing multiple RNA-seq samples by program loop or parallelism. Fourth, the priUTR program fits our computer systems (Scientific Linux release 6.10 Carbon), while other programs might not be readily compatible with our operational system.

Extensive bioinformatics efforts have been made to identify poly(A) sites or mRNA 3′ tails using RNA-seq data. These previous bioinformatic tools have been classified into four groups based on their principles of the algorithms (Chen et al., 2020). Some of the programs have been compared using benchmarking datasets from model species (*i.e.*, human, mouse, and *Arabidopsis*). The advantages and the limitations for each type of poly(A)-site prediction tools have been well discussed. Type 1 tools depend on a prior database of poly(A) sites in model animals or *Arabidopsis* and can effectively identify numerous poly(A) sites. Type 1 tools are limited to detect previously known poly(A) sites. Type 2 programs apply relatively mature software for transcriptome assembly, such as Scripture (Guttman et al., 2010), Cufflinks (Trapnell et al., 2012), and StringTie (Pertea et al., 2016). While type 2 programs could detect both 3′UTR extending and shortening events, they inherit the limitations of transcript assembly tools that are: (1) failure to distinguish the transcripts which overlap or reside closely between each other in the genome and (2) inability to distinguish a short transcript isoform embedded in longer ones from the same gene. Type 3 tools heavily depend on poly(A)-capped reads, which are scarce in regular RNA-seq data (Bayerlova et al., 2015; Kim et al., 2015), leading to low sensitivity of these methods. Type 4 programs detect sudden fluctuations of read density at the 3′ end of mRNAs in order to model 3′UTR usage. Type 4 programs tend to identify 3′UTR shortening events and have more inaccurate predictions caused by the heterogeneity of read coverage and non-biological variations (Szkop and Nobeli, 2017). Based on the information above, the type 2 methods which rely on transcript reconstruction and infer 3′UTR from assembled transcripts may be a potentially better choice for alternative 3′UTR study for non-model plant species with reference genome available.

As a type 2 bioinformatic tool for alternative 3′UTR prediction, priUTR has some limitations. It heavily relies on the accuracy of transcript annotation, and the accuracy of RNA-seq-based transcript reconstruction depends on the genic context. When multiple transcripts are produced from a gene model, of which the transcript with a shorter 3′UTR sequence embedded within a longer transcript, transcript reconstructions are prone to be inaccurate. Such cases have been identified in our study as the group 3 alternative 3′UTRs (**Figure 2**). In other situations, when several gene models are located closely in the same genomic locus, the prediction of 3′UTR extension tends to be false because the transcript could be mis-assembled to include neighboring genes. In this study, these likely mis-predicted alternative 3′UTRs were categorized as groups 1 and 2 (**Figure 2**). Another limitation of priUTR is that it discards those transcripts with less than 3 introns (including the intronless ones) since the transcript do not meet the requirement for partial intron-chain matching. This would limit the number of predicted alternative 3′UTRs. Besides that, priUTR is not applicable to non-model organisms without reference genomes. When priUTR is applied to an organism with a reference genome and poor annotations, caution should be taken as predicted alternative 3′UTRs could be previously unannotated 3′UTR regions.

Thanks to the identification of potential sources of mis-predicted alternative 3′UTRs, several improvements could be made to the priUTR tool in the near future. First, filters could be made to discard genes with multiple overlapped transcripts. Second, filters could be applied to discard those genes with neighbor genes located at 3′ downstream to avoid mis-assembly of transcript. Third, methods could be established to identify alternative 3′UTR transcripts with less than three introns.

Besides the application of priUTR method for characterizing alternative 3′UTRs in sorghum, we attempted to gain insights into the factors associated with the alternative 3′UTRs. Our results strongly indicate an association between m^{6}A modification and alternative 3′UTR (**Figure 3**). Recently, a maize study of RNA m^{6}A profiling reveals a clear indication between the m^{6}A modification and APA (Luo et al., 2020). There are both similarities and differences between the results of the maize m^{6}A study and the results of the present alternative 3′UTR study. In the maize study, genes were split into two sets, one set of genes with single poly(A) site and the other set of genes that contain at least two poly(A) sites and could have APA events. With such an approach for dividing the maize expressed genes, significant associations between potential APA genes and m^{6}A modified genes were observed (Luo et al., 2020). The maize study was a genome-wide, large-scale profiling of m^{6}A methylated mRNAs without actually determining the mRNAs with alternative 3′UTRs or poly(A) sites. In our sorghum study, the association between APA and m^{6}A methylation was discovered the other way around. We first identified the genes using alternative 3′UTRs and then discovered that many of the alternative 3′UTR genes could carry m^{6}A methylation. Another feature of our results is that a few hundreds of alternative 3′UTR genes were identified in sorghum, with the potential to carry m^{6}A modification. Together with enriched functions, our results could help to narrow down candidate genes for further functional studies with a focus on the cause-and-effect relationship between m^{6}A methylation and APA for specific functionally important genes. In addition, our results provide pieces of evidence that alternative 3′UTR usage for some genes (many functionally important) are genotype specific and could be inherent genetically.

For the m^{6}A hypothesis, public data from maize studies have been used (Luo et al., 2020; Miao et al., 2020). Maize and sorghum are close relatives, split just 11.9 million years ago (Mya) from a common progenitor (Swigonova et al., 2004). After the divergence, maize experienced allotetraploidization followed by diploidization (Gaut and Doebley, 1997; Xu and Messing, 2008). Therefore, the orthologous relationship between maize and sorghum genes is well established, and many gene regulations remain conserved (Zhang et al., 2017). Such conservation could help to establish an association between RNA m^{6}A modification and alternative 3′UTR. For the APA hypothesis, it remains possible that PA machinery proteins, stress responses, or developmental context might be involved in alternative 3′UTRs in sorghum. *Arabidopsis* and rice are plant model species divergent from sorghum. The authentic results of APA or alternative 3′UTRs related to stresses have only been reported by a handful of studies in plants (Sun et al., 2017; Téllez-Robledo et al., 2019). As far as our knowledge is concerned, the target genes of PA machinery proteins have only been studied in *Arabidopsis* (Thomas et al., 2012; Duc et al., 2013; Lin et al., 2017; Téllez-Robledo et al., 2019). Whether or to what extent these APA genes are evolutionarily conserved between *Arabidopsis* and sorghum remains to be determined. Thus, one of the possible explanations for the non-significant results between APA-associated genesets and the sorghum alternative 3′UTR genes could be that the APA genes that respond to stresses or are induced by mutations of PA proteins could be largely unconserved between sorghum and *Arabidopsis*. Overall, our results clearly lead to important testable hypotheses that: (1) m^{6}A modifications are functionally related to alternative 3′UTR/APA, yet the cause-and-effect relationship between them needs clarification, and (2) whether the genes with inter-genotype variations in alternative 3′UTR/APA could have functional consequences in sorghum; if yes, what are the consequences in molecular and phenotypic levels.

While conventional RNA-seq studies are prevalent in sorghum, studies on alternative 3′UTR/APA are scarce in non-model plant species. Previously, the full-length transcriptome of sorghum seedling was characterized by PacBio Iso-Seq and identified a few thousand transcripts with APA, demonstrating APA as a common phenomenon (Abdel-Ghany et al., 2016). We acknowledge that the priUTR-based 3′UTR analysis has limitations in the number of alternative 3′UTRs identified and the mis-predictions compared to those specialized pipelines based on the more advanced long-reads sequencing technologies (*i.e.*, PacBio and Oxford Nanopore) (Abdel-Ghany et al., 2016; Parker et al., 2020). These limitations are partly due to the short-read derived transcriptome assembly and to the limited ability in detecting transcripts with large variations using reference-guided mapping (Stark et al., 2019). Recently, new aspects of mRNA 3′ ends have been revealed to affect post-transcriptional regulation, including poly(A) length and sequence composition (Zhao et al., 2019). This suggests that a long-read-based method to integrate profiles of alternative 3′UTR/APA, poly(A) length and sequence, and m^6A modifications will be a powerful tool to understand 3′UTR-mediated regulation in the near future.

In summary, our work presents a transcriptome-wide profiling of alternative 3′UTRs in sorghum and identified hundreds of genes as the candidate genes to study the functional effects of alternative 3′UTR usage. The new insights reported here suggest future research directions for 3′UTR genotype specificity and the link between 3′UTR and epi-transcriptome modifications. Additionally, this study exemplifies alternative 3′UTR analysis using conventional RNA-seq, signifying 3′UTR analysis as a valuable addition to routine RNA-seq analysis in plants.

DATA AVAILABILITY STATEMENT

The RNA-seq data of sorghum used in this study can be found at NCBI Sequence Read Archive (SRA) under accession PRJNA413691. The priUTR program for 3′UTR analysis is available at github (https://github.com/mint1234/3UTR-). All data generated or analyzed during this study are included in this published article and its **Supplementary Material**.

AUTHOR CONTRIBUTIONS

MT and YL designed the study, performed the data analysis, and revised and finalized the manuscript. MT generated the data. YL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Waksman Genomics Facility and Waksman Institute of Microbiology.

ACKNOWLEDGMENTS

We thank Dr. Yaping Feng and Dr. Dibyendu Kumar at Waksman Institute of Microbiology, Rutgers University, for their contributions in the project of developing 3′ untranslated region analysis program, priUTR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.556749/full#supplementary-material

Supplementary Figure S1 | Schematic diagram of the priUTR algorithm.

Supplementary Figure S2 | Distribution of the expression levels for the five groups of predicted alternative 3′UTR genes.

Supplementary Figure S3 | Comparison of groups 1, 2, 3, and 5 of sorghum alternative 3′UTR genes with the APA- or m^6A- associated genesets in *Arabidopsis*, rice, and maize.

Supplementary Figure S4 | The distribution of the number of genes overlapped between the public genesets and the randomly selected genes in sorghum by using a permutation test.

Supplementary Figure S5 | The distribution of hypergeometric *p*-values between the public genesets and the randomly selected genes in sorghum by using a permutation test.

Supplementary Figure S6 | Genome browser views of the 23 genes with genotype-specific alternative 3′UTRs.

Supplementary Table S1 | Information about the 612 high-confidence alternative 3′UTR genes.

Supplementary Table S2 | Summary information about the public available genesets associated with m^6A and APA.

Supplementary Table S3 | Genesets for m^6A and APA.

Supplementary Table S4 | Hypergeometric tests for the overlap between sorghum alternative 3′UTR genes and the APA- or m^6A- associated genesets.

Supplementary Table S5 | Distribution of the expression levels of the high-confidence alternative 3′UTR genes in Rio, BTx406, and R9188, respectively.

Supplementary Table S6 | Functional enrichment results of group 4 genes and the m^6A-associated 3′UTR subset.

Supplementary Table S7 | Information about the 23 genes with genotype-specific alternative 3′UTR.

REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

Arefeen, A., Liu, J., Xiao, X., and Jiang, T. (2018). TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* 34, 2521–2529. doi: 10.1093/bioinformatics/bty110

Arribas-Hernández, L., Bressendorff, S., Hansen, M. H., Poulsen, C., Erdmann, S., and Brodersen, P. (2018). An m(6)A-YTH module controls developmental timing and morphogenesis in Arabidopsis. *Plant Cell* 30, 952–967. doi: 10.1105/tpc.17.00833

Bayerlova, M., Klemm, F., Kramer, F., Pukrop, T., Beibarth, T., Bleckmann, A., et al. (2015). Newly constructed network models of different WNT signaling cascades applied to breast cancer expression data. *PLoS One* 10:e0144014. doi: 10.1371/journal.pone.0144014

Bell, S. A., Shen, C., Brown, A., and Hunt, A. G. (2016). Experimental genome-wide determination of RNA polyadenylation in *Chlamydomonas reinhardtii*. *PLoS One* 11:e0146107. doi: 10.1371/journal.pone.0146107.g007

Birol, I., Raymond, A., Chiu, R., Nip, K. M., Jackman, S. D., Kreitzman, M., et al. (2015). Kleat: cleavage site analysis of transcriptomes. *Pac. Symp. Biocomput.* 2015, 347–358.

Bonfert, T., and Friedel, C. C. (2017). Prediction of poly(A) sites by poly(A) read mapping. *PLoS One* 12:e0170914. doi: 10.1371/journal.pone.0170914

Boyles, R. E., Brenton, Z. W., and Kresovich, S. (2019). Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *Plant J.* 97, 19–39. doi: 10.1111/tpj.14113

Brodersen, P., Sakvarelidze-Achard, L., Schaller, H., Khafif, M., Schott, G., Bendahmane, A., et al. (2012). Isoprenoid biosynthesis is required for miRNA function and affects membrane association of ARGONAUTE 1 in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1778–1783. doi: 10.1073/pnas.1112500109

Calvino, M., and Messing, J. (2012). Sweet sorghum as a model system for bioenergy crops. *Curr. Opn. Biotech.* 23, 323–329. doi: 10.1016/j.copbio.2011.12.002

Calvino, M., Bruggmann, R., and Messing, J. (2011). Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. *BMC Genomics* 12:356. doi: 10.1186/1471-2164-12-356

Chakrabarti, M., and Hunt, A. G. (2015). CPSF30 at the interface of alternative polyadenylation and cellular signaling in plants. *Biomolecules* 5, 1151–1168. doi: 10.3390/biom5021151

Chen, M., Ji, G., Fu, H., Lin, Q., Ye, C., Ye, W., et al. (2020). A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief. Bioinform.* 21, 1261–1276. doi: 10.1093/bib/bbz068

Cooper, E. A., Brenton, Z. W., Flinn, B. S., Jenkins, J., Shu, S., Flowers, D., et al. (2019). A new reference genome for Sorghum bicolor reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics* 20:420. doi: 10.1186/s12864-019-5734-x

Davidson, R. M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S. H., et al. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.* 71, 492–502.

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., et al. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.* 9:4844.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). AgriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70.

Duan, H. C., Wei, L. H., Zhang, C., Wang, Y., Chen, L., Lu, Z., et al. (2017). ALKBH10B is an RNA N6-methyladenosine demethylase affecting Arabidopsis floral transition. *Plant Cell* 29, 2995–3011. doi: 10.1105/tpc.16.00912

Duc, C., Sherstnev, A., Cole, C., Barton, G. J., and Simpson, G. G. (2013). Transcription termination and chimeric RNA formation controlled by Arabidopsis thaliana FPA. *PLoS Genet.* 9:e1003867. doi: 10.1186/1471-2164-12-514

Dugas, D. V., Monaco, M. K., Olson, A., Klein, R. R., Kumar, S., Ware, D., et al. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC Genomics* 12:514.

Fracasso, A., Trindade, L. M., and Amaducci, S. (2016). Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biol.* 16:115. doi: 10.1186/s12870-016-0800-x

Fu, H., Yang, D., Su, W., Ma, L., Shen, Y., Ji, G., et al. (2016). Genome-wide dynamics of alternative polyadenylation in rice. *Genome Res.* 26, 1753–1760. doi: 10.1101/gr.210757.116

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., et al. (2011). Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 21, 741–747. doi: 10.1101/gr.115295.110

Gaut, B. S., and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6809–6814. doi: 10.1073/pnas.94.13.6809

Gelli, M., Duo, Y., Konda, A. R., Zhang, C., Holding, D., and Dweikat, I. (2014). Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC Genomics* 15:179. doi: 10.1186/1471-2164-15-179

Gruber, A. J., Schmidt, R., Ghosh, S., Martin, G., Gruber, A. R., van Nimwegen, E., et al. (2018). Discovery of physiological and cancer-related regulators of 3’ UTR processing with KAPAC. *Genome Biol.* 19, 44.

Gruber, A. J., Schmidt, R., Gruber, A. R., Martin, G., Ghosh, S., Belmadani, M., et al. (2016). A comprehensive analysis of 3’ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* 26, 1145–1159. doi: 10.1101/gr.202432.115

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi: 10.1038/nbt.1633

Ha, K. C. H., Blencowe, B. J., and Morris, Q. (2018). QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.* 19:45.

Hafez, D., Ni, T., Mukherjee, S., Zhu, J., and Ohler, W. (2013). Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. *Bioinformatics.* 29, i108–i116.

Hennet, L., Berger, A., Trabanco, N., Ricciuti, E., Dufayard, J., Bocs, S., et al. (2020). Transcriptional regulation of sorghum stem composition: key players identified through co-expression gene network and comparative genomics analyses. *Front. Plant Sci.* 11:224. doi: 10.3389/fpls.2020.00224

Hong, L., Ye, C., Lin, J., Fu, H., Wu, X., and Li, Q. Q. (2018). Alternative polyadenylation is involved in auxin-based plant growth and development. *Plant J.* 93, 246–258. doi: 10.1111/tpj.13771

Hoque, M., Ji, Z., Zheng, D. H., Luo, W., Li, W., You, B., et al. (2013). Analysis of alternative cleavage and polyadenylation by 3’region extraction and deep sequencing. *Nat. Methods.* 10, 133–139. doi: 10.1038/nmeth.2288

Huang, Z., and Teeling, E. C. (2017). ExUTR: a novel pipeline for large-scale prediction of 3’-UTR sequences from NGS data. *BMC Genomics* 18:847.

International Wheat Genome Sequencing Consortium [IWGSC] (2008). (Shifting)the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:661.

Jan, C. H., Friedman, R. C., Ruby, J. G., and Bartel, D. P. (2010). Formation, regulation and evolution of Caenorhabditis elegans 3’ UTRs. *Nature* 469, 97–101. doi: 10.1038/nature09616

Ji, G., Guan, J., Zeng, Y., Li, Q. Q., and Wu, X. (2014). Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Brief. Bioinform.* 16, 304–313. doi: 10.1093/bib/bbu011

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527.

Johnson, S. M., Cummins, I., Lim, F. L., Slabas, A. R., and Knight, M. R. (2015). Transcriptomic analysis comparing stay-green and senescent Sorghum bicolor lines identifies a role for proline biosynthesis in the stay-green trait. *J Exp. Bot.* 66, 7061–7073. doi: 10.1093/jxb/erv405

Kebrom, T. H., McKinley, B., and Mullet, J. E. (2017). Dynamics of gene expression during development and expansion of vegetative stem internodes of bioenergy sorghum. *Biotechnol. Biofuels.* 10:159.

Kim, M., You, B. H., and Nam, J. W. (2015). Global estimation of the 3’ untranslated region landscape using RNA sequencing. *Methods* 83, 111–117. doi: 10.1016/j.ymeth.2015.04.011

Le Pera, L., Mazzapioda, M., and Tramontano, A. (2015). 3USS: a web server for detecting alternative 3’UTRs from RNA-seq experiments. *Bioinformatics* 31, 1845–1847. doi: 10.1093/bioinformatics/btv035

Lee, J. Y., Yeh, I., Park, J. Y., and Tian, B. (2007). PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucl. Acids Res.* 35, D165–D168.

Leiboff, S., and Hake, S. (2019). Reconstructing the transcriptional ontogeny of maize and sorghum supports an inverse hourglass model of inflorescence development. *Curr. Biol.* 20, 3410–3419. doi: 10.1016/j.cub.2019.08.044

Li, X., Nair, A., Wang, S., and Wang, L. (2015). “Quality Control of RNA-Seq Experiments,” in *RNA Bioinformatics in Methods in Molecular Biology*, Vol. 1269, ed. E. Picardi (New York, NY: Humana Press), 137–146. doi: 10.1007/978-1-4939-2291-8_8

Li, Y., Mehta, R., and Messing, J. (2018). A new high-throughput assay for determining soluble sugar in sorghum internode-extracted juice. *Planta* 248, 785–793. doi: 10.1007/s00425-018-2932-8

Li, Y., Tu, M., Feng, Y., Wang, W., and Messing, J. (2019b). Common metabolic networks contribute to carbon sink strength of sorghum internodes: implications for bioenergy improvement. *Biotechnol. Biofuels.* 12:274.

Li, Y., Wang, W., Feng, Y., Tu, M., Wittich, P. E., Bate, N. J., et al. (2019a). Transcriptome and metabolome reveal distinct carbon allocation patterns during internode sugar accumulation in different sorghum genotypes. *Plant Biotech. J.* 17, 472–487. doi: 10.1111/pbi.12991

Lin, J., Xu, R., Wu, X., Shen, Y., and Li, Q. Q. (2017). Role of cleavage and polyadenylation specificity factor 100: anchoring poly(A) sites and modulating transcription termination. *Plant J.* 91, 829–839. doi: 10.1111/tpj.13611

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* 10, 93–95. doi: 10.1038/nchembio.1432

Lorenzo, L., Sorenson, R., Bailey-Serres, J., and Hunt, A. G. (2017). Noncanonical alternative polyadenylation contributes to gene regulation in response to hypoxia. *Plant Cell* 29, 1262–1277. doi: 10.1105/tpc.16.00746

Lu, J., and Bushel, P. R. (2013). Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. *Gene* 527, 616–623. doi: 10.1016/j.gene.2013.06.052

Luo, G. Z., MacQueen, A., Zheng, G., Duan, H., Dore, L. C., Lu, Z., et al. (2014). Unique features of the m6A methylome in Arabidopsis thaliana. *Nat. Commun.* 5, 5630.

Luo, J. H., Wang, Y., Wang, M., Zhang, L. Y., Peng, H. R., Zhou, Y. Y., et al. (2020). Natural variation in RNA m6A methylation and its relationship with translational status. *Plant Physiol.* 182, 332–344. doi: 10.1104/pp.19.00987

Mandel, C. R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. *Cell Mol. Life Sci.* 65, 1099–1122. doi: 10.1007/s00018-007-7474-3

Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, M., et al. (2010). The landscape of C. *elegans* 3' UTRs. *Science* 329, 432–435.

Mathur, S., Umakanth, A. V., Tonapi, V. A., Sharma, R., and Sharma, M. (2017). Sweet sorghum as biofuel feedstock: recent advances and available resources. *Biotechnol. Biofuels.* 10:146.

McCarthy, D. J., Chen, Y., and Smyth, K. G. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucl. Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

McCormick, R. F., Truong, F. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

McKinley, B., Casto, A., Rooney, W., and Mullet, J. (2018). Developmental dynamics of stem starch accumulation in Sorghum bicolor. *Plant Dir.* 2, 1–15.

McKinley, B., Rooney, W., Wilkerson, C., and Mullet, J. (2016). Dynamics of biomass partitioning, stem gene expression, cell wall biosynthesis, and sucrose accumulation during development of Sorghum bicolor. *Plant J.* 88, 662–680. doi: 10.1111/tpj.13269

Meng, S., Peng, J. S., He, Y. N., Zhang, G. B., Yi, H. Y., Fu, Y. L., et al. (2015). Arabidopsis NRT1.5 mediates the suppression of nitrate starvation-induced leaf senescence by modulating foliar potassium level. *Mol. Plant.* 9, 461–470. doi: 10.1016/j.molp.2015.12.015

Miao, Z., Zhang, T., Qi, Y., Song, J., and Han, Z. (2020). Evolution of the RNA N6-methyladenosine methylome mediated by genomic duplication. *Plant Physiol.* 182, 345–360. doi: 10.1104/pp.19.00323

Mizuno, H., Kasuga, S., and Kawahigashi, H. (2016). The sorghum *SWEET* gene family: stem sucrose accumulation as revealed through transcriptome profiling. *Biotechnol. Biofuels* 9: 127.

Mizuno, H., Kasuga, S., and Kawahigashi, H. (2018). Root lodging is a physical stress that changes gene expression from sucrose accumulation to degradation in sorghum. *BMC Plant Biol.* 18:2. doi: 10.1186/s12870-017-1218-9

Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., et al. (2014). Energy Sorghum—a genetic model for the design of C4 grass bioenergy crops. *J. Exp. Bot.* 65, 3479–3489. doi: 10.1093/jxb/eru229

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., et al. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife.* 2020:e49658. doi: 10.7554/eLife.49658

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, H., Haberer, G., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551–556.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT. StringTie and Ballgown. *Nat. Prot.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Ping, X. L., Sun, B., Wang, L., Xiao, W., Yang, X., Wang, W., et al. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res.* 24, 177–189. doi: 10.1038/cr.2014.3

Preker, P. J., Lingner, J., Mivielle-Sebastia, L., and Keller, W. (1995). The FIP1 gene encodes a component of a yeast pre-mRNA polyadenylation factor that directly interacts with poly(A) polymerase. *Cell* 81, 379–389. doi: 10.1016/0092-8674(95)90391-7

Ritter, K., Chapman, S., Jordan, D., Godwin, I., and McIntyre, L. (2004). "Investigating the use of sweet sorghum as a model for sugar accumulation in sugarcane," in *Proceedings of the 4th International Crop Science Congress*, Brisbane.

Ruzicka, K., Zhang, M., Campilho, A., Bodi, Z., Kashif, M., Saleh, M., et al. (2017). Identification of factors required for m6A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. *New Phytol.* 215, 157–172. doi: 10.1111/nph.14586

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54:e6. doi: 10.1093/pcp/pcs183

Schrick, K., Fujioka, S., Takatsuto, S., Stierhof, Y. D., Stransky, H., Yoshida, S., et al. (2004). A link between sterol biosynthesis, the cell wall, and cellulose in Arabidopsis. *Plant J.* 38, 227–243. doi: 10.1111/j.1365-313x.2004.02039.x

Scutenaire, J., Deragon, J. M., Jean, V., Benhamed, M., Raynaud, C., Favory, J. J., et al. (2018). The YTH domain protein ECT2 is an m6A reader required for normal trichome branching in Arabidopsis. *Plant Cell* 30, 986–1005. doi: 10.1105/tpc.17.00854

Shen, Y., Venu, R. C., Nobuta, K., Wu, X., Notibala, V., Demirci, C., et al. (2011). Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. *Genome Res.* 21, 1478–1486. doi: 10.1101/gr.114744.110

Shenker, S., Miura, P., Sanfilippo, P., and Lai, E. C. (2015). IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA* 21, 14–27. doi: 10.1261/rna.046037.114

Shepard, P. J., Choi, E. A., Lu, J., Flanagan, L. A., Hertel, K. J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 17, 761–772. doi: 10.1261/rna.2581711

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R. III, et al. (2009). Molecular architecture of the human pre-mRNA 39 processing complex. *Mol. Cell.* 33, 365–376. doi: 10.1016/j.molcel.2008.12.028

Souter, M. A., Pullen, M. L., Topping, J. F., Zhang, X., and Lindsey, K. (2004). Rescue of defective auxin-mediated gene expression and root meristem function by inhibition of ethylene signalling in sterol biosynthesis mutants of Arabidopsis. *Planta* 219, 773–783.

Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z., and Zhu, J. K. (2018). UTR-dependent control of gene expression in plants. *Trends Plant Sci.* 23, 248–259. doi: 10.1016/j.tplants.2017.11.003

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2

Sui, N., Yang, Z., Liu, M., and Wang, B. (2015). Identification and transcriptomic profiling of genes involved in increasing sugar content during salt stress in sweet sorghum leaves. *BMC Genomics* 16:534.

Sun, H., Li, Y., Niu, Q., and Chua, N. H. (2017). Dehydration stress extends mRNA 3' untranslated regions with noncoding RNA functions in Arabidopsis. *Genome Res.* 27, 1427–1436. doi: 10.1101/gr.218669.116

Sun, X., Zheng, H., Li, J., Liu, J., Zhang, X., and Sui, N. (2020). Comparative transcriptome analysis reveals new lncRNAs responding to salt stress in sweet sorghum. *Front. Bioeng. Biotech.* 8:331. doi: 10.3389/fbioe.2020.00331

Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., et al. (2004). Close split of sorghum and maize genome progenitors. *Genome Res.* 14, 1916–1923. doi: 10.1101/gr.2332504

Szkop, K. J., and Nobeli, I. (2017). Untranslated parts of genes interpreted: making heads or tails of high-throughput transcriptomic data via computational methods: computational methods to discover and quantify isoforms with alternative untranslated regions. *Bioessays* 39:1700090. doi: 10.1002/bies.201700090

Téllez-Robledo, B., Manzano, C., Saez, A., Navarro-Neila, S., Silva-Navas, J., Lorenzo, L., et al. (2019). The polyadenylation factor FIP1 is important for plant development and root responses to abiotic stresses. *Plant J.* 99, 1203–1219. doi: 10.1111/tpj.14416

Thomas, P. E., Wu, X., Liu, M., Gaffney, B., Ji, G., Li, Q., et al. (2012). Genome-wide control of polyadenylation site choice by CPSF30 in Arabidopsis. *Plant Cell* 24, 4376–4388. doi: 10.1105/tpc.112.096107

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Prot.* 7, 562–579. doi: 10.1038/nprot.2012.016

Tu, M. (2020). *priUTR [Source Code]*. Available online at: https://github.com/mint1234/3UTR- (accessed April 15, 2020).

Turco, G. M., Kajala, K., Kunde-Ramamoorthy, G., Ngan, C. Y., Olson, A., Deshphande, S., et al. (2017). DNA methylation and gene expression regulation associated with vascularization in Sorghum bicolor. *New Phytol.* 214, 1213–1229. doi: 10.1111/nph.14448

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare omics data in plants: a case study in the crop species. *Maize*. *Plant Cell Environ.* 32, 1211–1229. doi: 10.1111/j.1365-3040.2009.01978.x

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., et al. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucl. Acid Res.* 46, D1190–D1196.

Varoquaux, N., Cole, B., Gao, C., Pierroz, G., Baker, C. R., Patel, D., et al. (2019). Transcriptomic analysis of field-droughted sorghum from seedling to maturity reveals biotic and metabolic responses. *Proc. Natl. Acad. Sci. U.S.A.* 116, 27124–27132. doi: 10.1073/pnas.1907500116

Wang, W., Wei, Z., and Li, H. (2014). A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics.* 30, 2162–2170. doi: 10.1093/bioinformatics/btu189

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., and Zhao, J. C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell. Biol.* 16, 191–198. doi: 10.1038/ncb2902

Wei, L., Song, P., Wang, Y., Lu, Z., Tang, Q., Yu, Q., et al. (2018). The m6A reader ECT2 controls trichome morphology by affecting mRNA stability in Arabidopsis. *Plant Cell* 30, 968–985. doi: 10.1105/tpc.17.00934

Wu, X., Gaffney, B., Hunt, A. G., and Li, Q. Q. (2014). Genome-wide determination of poly(A) sites in Medicago truncatula: evolutionary conservation of alternative poly(A) site choice. *BMC Genomics* 15:615. doi: 10.1186/1471-2164-15-615

Wu, X., Hu, W., Luo, H., Xia, Y., Zhao, Y., Wang, L., et al. (2016a). Transcriptome profiling of developmental leaf senescence in sorghum (Sorghum bicolor). *Plant Mol. Biol.* 92, 555–580. doi: 10.1007/s11103-016-0532-1

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q. Q., et al. (2011). Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12533–12538. doi: 10.1073/pnas.1019732108

Wu, X., Zhang, Y., and Li, Q. Q. (2016b). PlantAPA: a portal for visualization and analysis of alternative polyadenylation in plants. *Front. Plant Sci.* 7:889. doi: 10.3389/fpls.2016.00889

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., et al. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.* 5:5274.

Xie, Q., and Xu, Z. (2019). Sustainable agriculture: from sweet sorghum planting and ensiling to ruminant feeding. *Mol. Plant.* 12, 603–606. doi: 10.1016/j.molp.2019.04.001

Xu, J., and Messing, J. (2008). Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. *Proc. Natl. Acad. Sci. U.S.A.* 150, 14330–14335. doi: 10.1073/pnas.0807026105

Yang, Z., Li, J., Liu, L. U., Xie, Q., and Sui, N. (2020). Photosynthetic regulation under salt stress and salt-tolerance mechanism of sweet sorghum. *Front. Plant Sci.* 10:1722. doi: 10.3389/fpls.2019.01722

Yang, Z., Zheng, H., Wei, X., Song, J., Wang, B., and Sui, N. (2018). Transcriptome analysis of sweet Sorghum inbred lines differing in salt tolerance provides novel insights into salt exclusion by roots. *Plant Soil.* 430, 423–439. doi: 10.1007/s11104-018-3736-0

Ye, C., Long, Y., Ji, G., Li, Q. Q., and Wu, X. (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* 34, 1841–1849. doi: 10.1093/bioinformatics/bty029

Ye, C., Zhou, Q., Wu, X., Ji, G., and Li, Q. Q. (2019). Genome-wide alternative polyadenylation dynamics in response to biotic and abiotic stresses in rice. *Ecotoxicol. Environ.* 183:109485. doi: 10.1016/j.ecoenv.2019.109485

You, L., Wu, J., Feng, Y., Fu, Y., Guo, Y., and Long, L. (2014). APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucl. Acids Res.* 43, D59–D67.

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, F., Zhang, Y., Liao, J., Yu, Y., Zhou, Y., Feng, Y., et al. (2019). The subunit of RNA N6-methyladenosine methyltransferase OsFIP regulates early degeneration of microspores in rice. *PLoS Genet.* 15:e1008120. doi: 10.1371/journal.pgen.1008120

Zhang, J., and Wei, Z. (2016). An empirical Bayes change-point model for identifying 3′ and 5′ alternative splicing by next-generation RNA sequencing. *Bioinformatics.* 32, 1823–1831.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* 50, 1565–1573.

Zhang, Y., Ngu, D. W., Carvalho, D., Liang, Z., Qiu, Y., Roston, R. L., et al. (2017). Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* 29, 1938–1951. doi: 10.1105/tpc.17.00354

Zhang, L. M., Leng, C. Y., Luo, H., Wu, X. Y., Liu, Z. Q., Zhang, Y. M., et al. (2018). Sweet sorghum originated through selection of Dry, a plant-specific NAC transcription factor gene. *Plant Cell* 30, 2286–2307. doi: 10.1105/tpc.18.00313

Zhao, T., Huan, Q., Sun, J., Liu, C., Hou, X., Yu, X., et al. (2019). Impact of poly(A)-tail G-content on Arabidopsis PAB binding and their role in enhancing translational efficiency. *Genome Biol.* 20:189. doi: 10.1186/s13059-019-1799-8

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C. M., Li, C. J., et al. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell.* 49, 18–29. doi: 10.1016/j.molcel.2012.10.015

Zheng, Y., Drechsler, N., Rausch, C., and Kunze, R. (2016). The Arabidopsis nitrate transporter NPF7.3/*NRT*1.5 is involved in lateral root development under potassium deprivation. *Plant Signal. Behav.* 11:e1176819. doi: 10.1080/15592324.2016.1176819

Zhou, Q., Fu, H., Yang, D., Ye, C., Zhu, S., Lin, J., et al. (2019). Differential alternative polyadenylation contributes to the developmental divergence between two rice subspecies, japonica and indica. *Plant J.* 98, 260–276. doi: 10.1111/tpj.14209

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 28 October 2020
doi: 10.3389/fgene.2020.591194

Mapping Regulatory Determinants in Plants

Mary Galli[1]*, *Fan Feng*[1] *and Andrea Gallavotti*[1,2†]

[1]Waksman Institute of Microbiology, Rutgers University, Piscataway, NJ, United States, [2]Department of Plant Biology, Rutgers University, New Brunswick, NJ, United States

The domestication and improvement of many plant species have frequently involved modulation of transcriptional outputs and continue to offer much promise for targeted trait engineering. The cis-regulatory elements (CREs) controlling these trait-associated transcriptional variants however reside within non-coding regions that are currently poorly annotated in most plant species. This is particularly true in large crop genomes where regulatory regions constitute only a small fraction of the total genomic space. Furthermore, relatively little is known about how CREs function to modulate transcription in plants. Therefore understanding where regulatory regions are located within a genome, what genes they control, and how they are structured are important factors that could be used to guide both traditional and synthetic plant breeding efforts. Here, we describe classic examples of regulatory instances as well as recent advances in plant regulatory genomics. We highlight valuable molecular tools that are enabling large-scale identification of CREs and offering unprecedented insight into how genes are regulated in diverse plant species. We focus on chromatin environment, transcription factor (TF) binding, the role of transposable elements, and the association between regulatory regions and target genes.

Keywords: plant genomics, transcriptional regulation, chromatin, transcription factor binding, cis-regulatory regions

Edited by:
Wenqin Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Qi Li,
Center for Excellence in Molecular Plant Sciences (CAS), China
Yubing He,
Nanjing Agricultural University, China

Correspondence:
Mary Galli
marygalli@waksman.rutgers.edu

†ORCID:
Andrea Gallavotti
orcid.org/0000-0002-1901-2971

Specialty section:
This article was submitted to Evolutionary and Population Genetics,
a section of the journal Frontiers in Genetics

Received: *04 August 2020*
Accepted: *28 September 2020*
Published: *28 October 2020*

Citation:
Galli M, Feng F and Gallavotti A (2020) Mapping Regulatory Determinants in Plants. Front. Genet. 11:591194. doi: 10.3389/fgene.2020.591194

REGULATORY REGIONS AND MECHANISMS REVEALED BY CLASSIC STUDIES

Mining trait-associated genetic factors has traditionally been performed using classical genetics, GWAS, and QTL analysis. Examples from these studies serve as excellent guides for understanding the molecular basis of phenotypic diversity (Deplancke et al., 2016). In particular, the regions corresponding to several beneficial traits associated with the domestication and diversification of many plant species from their wild relatives have been mapped by these approaches and frequently shown to be located in the intergenic space, sometimes residing up to 100 kb from the closest protein coding genes (**Figure 1A**; Olsen and Wendel, 2013; Rodgers-Melnick et al., 2016; Swinnen et al., 2016; Lu et al., 2019). Correspondingly, these traits involve variations in gene expression, with variants affecting either the level of expression or the spatial and/or temporal pattern of expression of certain genes (**Figure 1B**; Meyer and Purugganan, 2013; Springer et al., 2019). Unlike changes to protein-coding genes which often result in easily interpretable loss-of-function alleles, the exact causative features underlying functional cis-regulatory

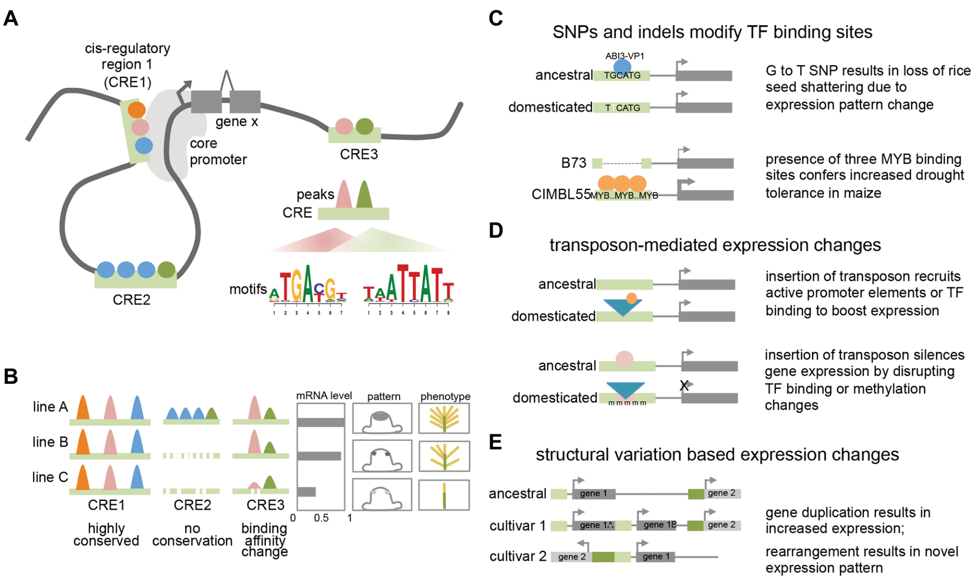

A
cis-regulatory region 1 (CRE1)
gene x
core promoter
CRE3
peaks
CRE
CRE2
motifs
B
line A
line B
line C
CRE1
highly conserved
CRE2
no conservation
CRE3
binding affinity change
mRNA level
0
0.5
1
pattern
phenotype
C
SNPs and indels modify TF binding sites
ABI3-VP1
ancestral
TGCATG
domesticated
T CATG
G to T SNP results in loss of rice seed shattering due to expression pattern change
B73
CIMBL55
MYB..MYB..MYB
presence of three MYB binding sites confers increased drought tolerance in maize
D
transposon-mediated expression changes
ancestral
domesticated
insertion of transposon recruits active promoter elements or TF binding to boost expression
ancestral
domesticated
mmmmm
X
insertion of transposon silences gene expression by disrupting TF binding or methylation changes
E
structural variation based expression changes
ancestral
gene 1
gene 2
cultivar 1
gene 1A
gene 1B
gene 2
cultivar 2
gene 2
gene 1
gene duplication results in increased expression;
rearrangement results in novel expression pattern

FIGURE 1 | Plant transcriptional regulation **(A)** model of plant transcriptional regulation at gene X. Colored circles represent different TFs binding to three distinct cis-regulatory regions (CREs; light green bars) that can contact the core promoter *via* DNA looping. Motifs enriched within binding peaks for two TFs are shown for CRE3. **(B)** Conservation and variation of TF binding events among different lines or accessions. Colored peaks represent different TF binding events within CREs. mRNA expression levels, cell-type specific expression pattern, and resulting phenotype are shown. **(C)** Examples showing how single nucleotide polymorphisms (SNPs) and indels can result in expression and phenotypic changes. **(D)** Examples showing how transposon insertions can result in expression and phenotypic changes. **(E)** Examples showing how structural variants can result in expression changes.

regions (CREs) are currently difficult to identify given the variable nature of regulatory elements, their frequent gene-distal location, and the lack of an obvious rigid code that determines their functionality. Understanding the molecular nature of these changes however lies at the heart of our ability to accelerate crop improvement using CRISPR-based targeted engineering of useful traits and traditional breeding (Rodríguez-Leal et al., 2017; Chen et al., 2019; Eshed and Lippman, 2019; Springer et al., 2019).

In several cases, the molecular nature of the phenotypic variation has been determined and found to be associated with a range of different causes. These include single nucleotide polymorphisms (SNPs) that affect transcription factor (TF) binding, either by disrupting or recruiting additional TF binding sites. For example, a G to T nucleotide change located 12 kb upstream of the *qSH1* gene in rice, a BEL-type homeobox TF, is believed to disrupt an ABI3-VP1 TF binding site (Konishi et al., 2006). This results in a loss of *qSH1* expression in the pedicel abscission zone and a subsequent non-shattering phenotype that facilitated higher harvesting yields. Alternatively, changes in TF binding can also involve advantageous gain of function elements. A GWAS screen for drought tolerance in maize identified a 366 bp region located in the proximal upstream region of *ZmVPP1*, a vacuolar-type H$^+$-pyrophosphatase, that conferred increased drought tolerance in several varieties (Wang et al., 2016). This fragment contains three putative MYB binding sites, which were shown to increase expression of *ZmVPP1* relative to the drought-sensitive maize line B73, which lacks the MYB binding sites.

In other cases, functional traits associated with cis-regulatory elements (CREs) may not involve nucleotide variations that directly correspond to known TF binding sites but are instead located nearby. This is the case for the rice *GW7* gene, which affects grain width and grain quality (Wang et al., 2015b). Certain rice varieties were found to contain two short indels directly adjacent an SBP16/GW8 TF binding motif in the proximal upstream region of *GW7*. These indels do not directly disrupt the TF binding motif but do appear to lower expression of *GW7* relative to varieties in which the indels are absent. Given that regulatory regions typically contain multiple different TF binding sites (Hardison and Taylor, 2012; Ricci et al., 2019), such examples could indicate that these divergent regions simply correspond to unknown TF binding sites and reflect the incompleteness of TF binding motif characterization in plants. Alternatively, they could alter local DNA shape (i.e., the sequence-dependent DNA structure surrounding the motif) or spacing between adjacent motifs, among other factors that contribute to the complexity of TF binding specificity (Slattery et al., 2014). Such examples highlight the need for comprehensive annotation of TFs and other regulatory regions. Similar examples have been noted in non-plant studies, where there is accumulating evidence that causative SNPs frequently do not directly affect TF binding motifs, but may impact cooperative or collaborative binding of TF complexes (Deplancke et al., 2016).

Transposon insertions in regulatory regions can also influence gene expression of adjacent genes, resulting in either elevated or suppressed gene expression levels, and likely act through a variety of mechanisms (Hirsch and Springer, 2017; Zhao et al., 2018).

A classic example of the former in plants is the presence of a *Hopscotch* element located ~60 kb upstream of the *TEOSINTE BRANCHED1 (TB1)* gene, a TCP-family TF that determines the apical dominance of domesticated maize relative to its highly branched wild ancestor teosinte (Studer et al., 2011). The *Hopscotch* element enhances the expression of *TB1* through an unknown mechanism. Interestingly, a nearby *Tourist* transposon within the same enhancer appears to repress expression of *TB1*, highlighting the dynamic nature of transcriptional changes conferred by transposable elements. Another illustrative example includes the insertion of a *Copia* retroelement in the proximal upstream region of the *RUBY* gene in blood oranges. *RUBY* encodes a MYB TF involved in anthocyanin production and its expression level is increased by cold-induced expression conferred by sequences within the long terminal repeat (LTR) that are hypothesized to harbor either promoter-like features with a TATA box and TSS, or other upstream activating sequences (Butelli et al., 2012). These examples suggest that like other cases from animals, transposons may act as novel promoters by recruiting the basal transcriptional machinery or introducing tissue-specific TF binding sites (or disrupting repressive TF binding sites; Butelli et al., 2012; Sundaram et al., 2014).

Transposon insertions within regulatory regions are also able to negatively impact gene expression. They can do this by disrupting existing TF binding sites or other regulatory features, or *via* epigenetic changes typically involving repressive DNA methylation (Huang and Ecker, 2018). For example, one of the major factors determining fruit color in grape species, is caused by a *Gypsy-like* retrotransposon insertion, *Gret1*, in the upstream region of *MYBA1*, involved in berry anthocyanin production. As opposed to the *RUBY* blood orange case described earlier, the presence of *Gret1* results in loss of gene expression and the white-colored berries typical of chardonnay (Kobayashi et al., 2004). Similar cases of transposon mediated gene repression are also seen in maize at the *ZmCCT10* and *ZmCCT9* loci, two genes involved in flowering-time regulation whose causative transposon insertions reside 2.5 and 57 kb upstream, respectively (Yang et al., 2013; Huang et al., 2017). In general, the mechanisms of how such transposon associated CREs influence expression are not fully understood although these examples and others suggest they can affect both distal enhancers and proximal regulatory regions. In other cases involving transposon insertions in regulatory regions, changes in DNA methylation have been documented as the underlying cause of stable gene downregulation (Hirsch and Springer, 2017). Examples of such epialleles include a methylated *hAT* element inserted in the proximal regulatory region of the melon *CmWIP* gene, which controls sex determination (Martin et al., 2009) and a SINE retrotransposon inserted upstream of the tomato *VTE3* gene, involved in vitamin E biosynthesis (Rossi et al., 2014). Possible mechanisms that explain stable transposon-triggered repression include spreading of methylation marks from the TE into the adjacent regulatory region, thus altering chromatin accessibility or blocking TF motif binding (many TFs preferentially bind unmethylated sites; Eichten et al., 2012; O'Malley et al., 2016; Huang et al., 2018). Overall, these examples as well as studies analyzing global transposon location (i.e., 86% of maize genes contain a TE within 1 kb of the gene; Hirsch and Springer, 2017) and association with eQTL, suggest that TE-driven transcriptional influence is frequent and in certain genomes may be major drivers of regulatory variation (Zhao et al., 2018; Noshay et al., 2020).

Although far less frequent than regulatory changes associated with TE insertions, there are several reports of regulatory epialleles that appear to have formed spontaneously. These include the *Colorless non-ripening (Cnr)* mutant allele of tomato, which encodes an SBP TF that affects color ripening (Manning et al., 2006). In the *Cnr* mutant, the upstream regulatory region of the *Cnr* gene is stably hypermethylated throughout development, leading to reduced expression of the gene (Zhong et al., 2013). Interestingly, the methylated sites are adjacent to two MADS-box TF binding sites bound by RIPENING INHIBITOR1 (RIN1; a MADS-box TF) in ChIP-seq (Zhong et al., 2013) suggesting that methylation changes in the *Cnr* epimutant could impact TF binding.

Finally, structural variants have also been shown to affect regulatory outputs by altering gene copy number and/or the arrangement or composition of CREs (Alonge et al., 2020), highlighting the modular architecture of regulatory elements. In the case of inversions, a certain gene may become located adjacent to an otherwise distally located gene or regulatory region and assume novel expression patterns. This appears to be the case for the classic *Tunicate* allele of maize, which shows unusually long glumes in both inflorescences as a result of ectopic expression from the 3' region of a gene normally located 1.8 Mb away (Han et al., 2012). Other structural variants include segmental duplications that increase gene copy number. While these do not directly involve changes in CREs, they do appear to be a subtle but possibly frequent mechanism of trait-associated transcriptional modulation in certain species (Alonge et al., 2020). Other situations in which putative regulatory regions are rearranged or duplicated are less clear. A good example of this is the ~4 kb DICE distal enhancer element in maize which confers increased expression of the *BX1* gene and consequently increased herbivore resistance (Betsiashvili et al., 2015; Zheng et al., 2015). The DICE element appears to be a divergent duplication of nearby sequences, and the increased expression may result from increased recruitment of specific TFs (Galli et al., 2018). Additional examples from maize include the classic cases of the *b1* and *Vgt1* loci, both of which are associated with structural variation in distal non-coding regions that results in epigenetic changes (Stam et al., 2002; Castelletti et al., 2014).

Detailed genetic and molecular characterization of QTL and classic cases have established a solid groundwork for understanding how regulatory changes influence many phenotypic traits in plants. However, they likely represent only a small fraction of the genetic variation and molecular mechanisms that govern transcriptional response for quantitative traits. Recently-developed genomics based techniques are paving the way for large-scale mining of putative CREs and begin to outline certain molecular signatures that correlate with gene expression and are conserved across species and accessions

(Maher et al., 2018; Lu et al., 2019; Alonge et al., 2020). Ultimately, combining both genetic and genome-wide studies will prove a powerful technique to better understand beneficial traits.

GENOME-WIDE IDENTIFICATION OF cis-REGULATORY REGIONS

Regulatory DNA in eukaryotes is generally characterized by chromatin accessibility, low DNA methylation, and is often associated with distinct histone modifications (Marand et al., 2017; Oka et al., 2017; Klemm et al., 2019; Lu et al., 2019). In plants, several recent studies have taken advantage of these properties to mine candidate regulatory elements at the genomic level (Sullivan et al., 2014; Rodgers-Melnick et al., 2016; Oka et al., 2017; Lü et al., 2018; Maher et al., 2018; Lu et al., 2019; Ricci et al., 2019; Parvathaneni et al., 2020). Such approaches are critical because while previous promoter and QTL studies suggest that most regulatory elements appear to lie within 1–2 kb upstream of the gene body in smaller genomes such as *Arabidopsis*, in larger genomes, regulatory regions reside within a much broader upstream area, with distal elements occasionally located hundreds of kb from the genes they regulate, making their identification by traditional means arduous. Therefore, the identification of accessible chromatin regions (ACRs) using techniques such as ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), MNaseHS (micrococcal nuclease hypersensitivity), and DNaseHS (DNAse hypersensitivity) has been highly informative for mapping regulatory regions in plants, revealing their frequency, size, and location, as well as many other important aspects. These studies demonstrate that ACRs are most often found near transcription start and end sites, but can also frequently be found over 2–200 kb from any gene depending on the species (Sullivan et al., 2014; Rodgers-Melnick et al., 2016; Oka et al., 2017; Maher et al., 2018; Lu et al., 2019; Ricci et al., 2019). They also show that ACRs can be condition and tissue-specific, highlighting the dynamic nature of chromatin (Sullivan et al., 2014; Rodgers-Melnick et al., 2016; Oka et al., 2017; Maher et al., 2018; Ricci et al., 2019; Parvathaneni et al., 2020). In support of their functionality, most identified ACRs are enriched for TF binding events and motifs and show transcriptional enhancer activity (see below for more detail; Sullivan et al., 2014; Ricci et al., 2019). Importantly, it was shown that SNPs in ACRs explain up to 40% of the variability in quantitative traits in maize and in particular overlap with several classically defined distal QTL discussed previously, substantiating their functionality and highlighting the role of regulatory regions in modulating phenotypes (Rodgers-Melnick et al., 2016; Ricci et al., 2019).

A landmark, cross-species comparative study of 13 angiosperm species with genome sizes ranging from ~100 to 5,000 Mb demonstrated that ACRs account for 0.2–6.5% of the total genome of a species and that their location varies according to genome size (Lu et al., 2019). For example, while the total sequence length of ACRs was fairly consistent across species regardless of genome size, large genomes showed a greater percentage of distally located ACRs (i.e., small genomes such as *Arabidopsis* showed that only ~6% of all ACRs were distal compared to ~46% in barley). Transposon insertions were found to be one of the main factors contributing to this occurrence, presumably pushing ACRs away from genes (Lu et al., 2019). Transposons themselves also appeared to be responsible for creating certain species-specific distal ACRs, as noted previously from classical studies (see above i.e., maize *TB1*). The controlled parallel nature of the Lu et al. (2019) study also allowed several important cross-species observations such as the finding that the number of ACRs correlated with the number of genes within a species and that many distal ACRs were conserved between sister species. Overall, an important finding from this study is that large and small plant genomes appear to be structured differently, despite harboring many of the same genetic pathways and gene regulatory networks (Lu et al., 2019). This underscores the importance of empirically mining sufficient amounts of regulatory information both for direct application in a species of interest such that ultimately such information will enable accurate machine learning predictions in other crop species.

A major factor in the characterization of putative regulatory regions is determining their functionality. In animals, regulatory regions are generally categorized into classes such enhancers, insulators, or promoters depending on their role in gene expression (Andersson and Sandelin, 2020). These terms however remain somewhat ambiguous despite an enormous effort toward their classification, perhaps because the elements themselves are heterogeneous (Andersson and Sandelin, 2020; Gasperini et al., 2020). In plants, these operational definitions are even more vague; however, studies have begun to tease out some common trends. Plant adapted versions of massively parallel promoter and enhancer reporter assays such as self-transcribing active regulatory region sequencing (STARR-seq; Ricci et al., 2019; Jores et al., 2020), show that many ACRs are capable of enhancing gene expression (Ricci et al., 2019). Traditional STARR-seq works by inserting fragments either from randomly sheared genomic sequence, BAC libraries, or small fragments such as those from ATAC-seq and placing them downstream of a cassette containing a minimal promoter fused to GFP (Arnold et al., 2013). Because enhancers are assumed to be capable of controlling gene expression regardless of distance or orientation (according to the classical definition), STARR-seq allows for self-driven transcription of the element and quantitative readout. In maize, both proximal and distal ACRs were found to show a general enhancement of activity, relative to randomly selected regions with similar features (Ricci et al., 2019). On the other hand, a modified version of STARR-seq using transient transfection in tobacco leaves found that four known plant enhancers gave the strongest transcriptional output when placed immediately upstream of a minimal promoter and were not active when placed in the 3'UTR of the reporter gene (Jores et al., 2020). Further studies are needed to tease out the functional determinants and optimal architecture of the various classes of regulatory elements. Given their utility to generate synthetic transcriptional units for agricultural improvement (Liu and Stewart, 2016), findings from such assays, and approaches could be directly applicable in plants, unlike in animals.

Genomes also typically harbor specific chromatin features that serve as another potential source of regulatory information (Marand et al., 2017). In animals, ACRs are often associated with distinct histone modifications that correlate with gene expression outputs (Hardison and Taylor, 2012; Gasperini et al., 2020). There has been much focus placed on using unique signatures of these various chromatin marks to identify particular classes of regulatory elements (e.g., enhancers) to aid genome annotation efforts and understand how chromatin environment impacts gene expression. However, it is widely accepted that operational definitions based on these biochemical marks serve as a guide rather than a fixed rule (Gasperini et al., 2020). Several large-scale studies have profiled histone modifications in various plant species (Oka et al., 2017; Lü et al., 2018; Lu et al., 2019; Peng et al., 2019; Ricci et al., 2019), and detailed analysis suggests that as in animals, certain chromatin signatures correlate with gene expression levels: expressed genes are enriched for H3K4me3, H3K56ac, and H2A.Z at the transcription start site, whereas repressed genes are enriched for H3K27me3 and H2A.Z (Lu et al., 2019). Furthermore, in maize, it appears that H3K27me3 marks often correspond to tissue-specific genes while H3K4me1 and H3K4me3 tend to mark broadly expressed genes (Lu et al., 2019; Peng et al., 2019; Ricci et al., 2019). Combining histone modification data with ACRs found that H3K9/K27/K56ac marks were generally associated with high expression levels of nearby genes and may represent enhancers. Distal ACRs instead marked by H3K27me3 tended to be located near genes with lower levels of expression and may represent repressor elements. Interestingly, it appears that some plant histone modification trends differ from those found in animals (Lu et al., 2019). For example, while H3K4me1 marks are typically found at distal CREs, in plants, this modification was not frequently associated with distal CREs (Lu et al., 2019).

Finally, DNA methylation maps are also highly valuable for mining regulatory information (Crisp et al., 2019). Prior studies have noted that most ACRs are hypomethylated, and in large genomes that are typically heavily methylated, unmethylated regions (UMRs) serve as an excellent tool to mine functional CREs (Crisp et al., 2020). Importantly, UMRs tend be static across most tissues and conditions in plants, whereas ACRs and histone modifications are often dynamic. Therefore, UMRs from a single tissue can be used to locate CREs, and when paired with chromatin accessibility data from a dissimilar tissue, can reveal CREs potentially set to become accessible or expressed in another tissue (Crisp et al., 2020).

Overall, these various genome-wide approaches for mining regulatory elements are generating highly informative maps that are crucial for understanding regulatory dynamics (**Figure 2**). Such data are critical for locating regulatory regions for use in transgenic studies or harnessing tissue-specific promoters for genetic engineering purposes.

TRANSCRIPTION FACTORS: DRIVERS OF GENE EXPRESSION

At the heart of transcriptional regulation is DNA-binding TFs and TF complexes bound to CREs. Transcription factors recognize short DNA sequence motifs in regulatory regions of their target genes and control the gene expression changes responsible for plant developmental programs and environmental responses. TFs bind to family-specific DNA motifs that contain four to six nucleotides, although many instances of longer and more

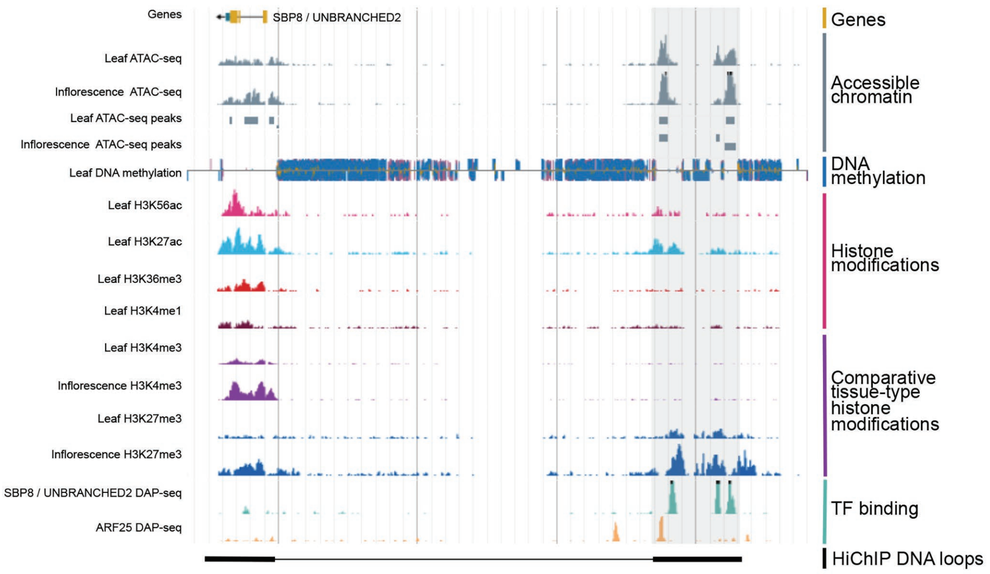


FIGURE 2 | Integration of various types of genomic regulatory data allows for the identification of CREs. Shown is a genome browser view of putative distal CRE (gray shaded region) located 40 kb upstream of the SBP8/UNBRANCHED2 gene in maize. Data obtained from Ricci et al., 2019.

complex architecture are known (Jolma et al., 2013; Weirauch et al., 2014; O'Malley et al., 2016). Particularly, in the case of short motifs, it is clear that TFs do not bind to all instances of these motifs within a given genome, suggesting that other factors also influence binding specificity (Hardison and Taylor, 2012; Todeschini et al., 2014). These have been shown to include DNA shape, i.e., the DNA sequence surrounding the motif, which is not directly bound by the TF (Slattery et al., 2014), as well as other factors such as the presence of proximally located motifs that can be bound by cooperating TFs (Deplancke et al., 2016). However, while these features play a role, the precise determinants of TF binding specificity remain unclear. One of the many additional interesting features of TF binding is the tendency for diverse TFs to bind in clusters, often lying within a region of open chromatin (**Figure 1**; Gasperini et al., 2020). This has been observed in many animal systems where a large number of genome-wide TF binding maps are available, and appears to occur in plants as well (see below for more detail). It remains unclear how these clusters of TFs are involved in gene regulation; however, the modular/combinatorial binding nature of these regulatory regions (i.e., multiple TFs binding) appears to allow genes to be controlled in tissue-specific or temporal manner (Spitz and Furlong, 2012). In plants, this is particularly intriguing from an agronomic engineering perspective because it suggests that phenotypes associated with distinct organs (i.e., ear traits but not tassel traits in maize) could be separated, allowing specific alterations to one organ or conditional response without altering another with a less desirable phenotype (Dong et al., 2019).

There are several methods by which to identify TF binding. ChIP-seq is the current gold-standard method for determining *in vivo* binding sites of TFs in live cells (Johnson et al., 2007). This method enables the identification of genomic binding sites in a tissue-specific chromatin context with high resolution (Park, 2009; Kaufmann et al., 2010). DNA-protein complexes are immunoprecipitated using an antibody specific to the protein of interest or a tag that is fused to the protein, and DNA is purified from the immunoprecipitated complex and subjected to next-generation sequencing. Several key factors that contribute to high-quality data in ChIP-seq, include antibody selection, negative controls, and biological replicates (Park, 2009; Kidder et al., 2011; Landt et al., 2012). Because of its *in vivo* context, ChIP-seq captures DNA bound both directly and indirectly by the TF of interest. This can include sites bound by hetero- or multimeric complexes. Many small and medium scale ChIP-seq studies have been carried out in *Arabidopsis* in contrast to the handful that have been performed in larger genomes such maize and soybean (Bolduc et al., 2012; Huang et al., 2012; Gregis et al., 2013; Lau et al., 2014; Tsuda et al., 2014; Li et al., 2015; Pautler et al., 2015; Jung et al., 2016; Song et al., 2016; Feng et al., 2018; Jo et al., 2020). A major limitation to ChIP-seq in plants is the time and effort required to either create transgenic lines or generate antibodies.

Performing ChIP-seq using protoplasts that transiently express epitope-tagged transcription factors is an alternative approach (Kong et al., 2012; Lee et al., 2017; Tu et al., 2020), as in some cases, specific antibodies against an endogenous protein of interest or transgenic lines expressing the protein of interest fused with a tag in a mutant background are unavailable. Protoplasts can be obtained either from mesophyll or other tissues such as root or stem and are transformed with a plasmid that expresses the protein of interest fused with an epitope-tag driven by a ubiquitously expression promoter such as 35S (Hernandez et al., 2007; Yoo et al., 2007; Kong et al., 2012; Para et al., 2014). ChIP-seq using protoplasts has obvious advantages as it bypasses the requirements for antibody or transgenic plants; however, overexpression of proteins in protoplasts might lead to altered genomic binding profiles due to excess protein in the cell (Kidder et al., 2011). A recent large-scale study using this approach in maize to map the binding sites of 104 TFs in leaves observed several key findings. As seen in animals, plant TF binding sites clustered together, covering ~2% of the maize genome and reinforcing the emerging paradigm that multiple TFs are needed for regulation of a single locus (Tu et al., 2020). These results also suggest co-binding appears to be important for TF specificity in maize (Tu et al., 2020).

Another modified version of ChIP-seq is cleavage under targets and release using nuclease (CUT&RUN), a chromatin profiling strategy in which antibody-targeted controlled cleavage by micrococcal nuclease releases specific protein-DNA complexes into the supernatant for paired-end DNA sequencing (Skene and Henikoff, 2017; Skene et al., 2018). Compared to ChIP-seq, CUT&RUN has several key advantages such as no crosslinking, which avoids false positive signals; *in situ* targeted digestion, which greatly reduces background; efficiency, as it can be finished in a day; and high signal-to-noise ratio, requiring only one tenth of the sequencing depth as ChIP-seq.

DAP-seq is an *in vitro* alternative to ChIP-seq (O'Malley et al., 2016). DAP-seq works by combining a standard Illumina-based genomic DNA sequencing library together with an *in vitro* expressed affinity-tagged TF coupled to magnetic beads. After a series of washes, TF-bound DNA is eluted, enriched, and barcoded for multiplexing, followed by next-gen sequencing (Bartlett et al., 2017). Resulting reads produce genome-wide peak maps similar to ChIP-seq, but often with higher resolution. A main advantage of DAP-seq is that it combines the low cost and high throughput of an *in vitro* assay with DNA in its native sequence context thereby preserving DNA structure and DNA methylation marks that are known to impact TF binding (O'Malley et al., 2016). Bound fragments are directly mapped to a genome unlike other *in vitro* assays such as HT-SELEX and protein binding microarrays, which report only motifs (Jolma et al., 2013; Weirauch et al., 2014). DAP-seq has been used to generate high quality peak maps for 529 *Arabidopsis* TFs and several maize TFs (O'Malley et al., 2016; Galli et al., 2018; Ricci et al., 2019). This data revealed many informative properties of plant TFs such as high frequency at which TFs from the same family- or subfamily-type bind similar sites, that TFs bind a very small fraction of all motif instances, and again that TFs cluster together in proximal promoters (and distal enhancers which are often located over 20–100 kb from their putative target gene in maize). Comparative studies of DAP-seq showed significant overlap with ChIP-seq data;

however, DAP-seq generally produces more peaks than ChIP-seq suggesting that DAP-seq captures binding events that take place independent of tissue- or condition-specific chromatin information (O'Malley et al., 2016).

Genome wide TF binding maps generated by these various techniques will be essential for understanding factors influencing both TF binding and TF activity. Yet while TFs are the major modulators of transcriptional activity, and their individual importance is often evident from mutations with severe developmental consequences, how TFs actually modulate gene expression remains largely unclear (de Boer et al., 2020). As in animal systems, it is also clear that not all TF binding is functional (Spitz and Furlong, 2012; Para et al., 2014; Brooks et al., 2019; Gasperini et al., 2020). Therefore, another challenge will be establishing determinants of TF activity and how these are influenced by factors such as position of binding sites, binding site strand, helical position, and protein interactions (de Boer et al., 2020). As mentioned previously, TF binding sites often cluster together and form cis-regulatory modules (CRMs; Hardison and Taylor, 2012) which themselves could impact TF activity. These CRMs and the individual TF binding sites within are often conserved within and across species indicating that together they may be important for TF activity and gene expression. Deciphering the degree to which plant TFs may work cooperatively will require dissection of CRMs using both natural variation and targeted genomic editing to better understand these regulatory regions.

INTERACTIONS BETWEEN REGULATORY REGIONS AND GENES: TARGET GENE IDENTIFICATION AND FUNCTIONAL CONSEQUENCES OF 3D CONFORMATION

An essential aspect of mining regulatory elements in any genome is being able to associate a putative regulatory region with a target gene or genes, and its expression dynamics. This remains a particularly challenging task in large genomes where regulatory regions may be located hundreds of kb away (Pliner et al., 2018). The current model of regulatory region-gene interactions involves looping of DNA in 3D space to allow physically distant regions to contact core promoters (**Figure 1A**; Shlyueva et al., 2014), and until recently this general eukaryotic model was derived largely from data in animals. Several plant studies using chromosome conformation capture (3C)-based techniques such as Hi-C and other variants, which capture global chromatin interactions (van Steensel and Dekker, 2010), have now shown that plant 3D chromatin organization generally resembles that reported in animals (Wang et al., 2015a, 2017; Dong et al., 2017; Liu et al., 2017; Mascher et al., 2017; Li et al., 2019; Peng et al., 2019; Ricci et al., 2019; Sun et al., 2020), despite the absence of certain proteins such as CTCF that are associated with this phenomenon in animals (Liu et al., 2017; Rowley et al., 2017). In these assays, chromatin contacts within a particular tissue are first cross-linked with formaldehyde, sheared to linearize the DNA, and then DNA ends are ligated together. The resulting ligated DNA is sequenced and consists of fragments that may not reside close in linear genomic space but are contacted in 3D space, often reflecting long-range spatial associations. Importantly, comparison among various plant genomes suggests that the 3D architecture of small, compact plant genomes such as *Arabidopsis* which tend to have CREs located within or near genes, differs from that of larger plant genomes which often form extensive long-range chromatin loops (Wang et al., 2015a, 2017; Dong et al., 2017; Liu et al., 2017; Ricci et al., 2019).

Bulk chromatin capture techniques such as Hi-C are often limited in their resolution, preventing the detailed empirical mapping of linkages between regulatory regions and target genes, and thus limiting the functional mapping of regulatory elements. More focused techniques such as Hi-ChIP and ChIA-PET use antibodies to enrich for a specific subset of chromatin interactions that are associated with RNA polymerase II, a particular histone modification, or transcription factor, offering greater resolution at a lower sequencing depth (Fullwood et al., 2009; Mumbach et al., 2016). A series of reports that mapped 3D chromatin interactions using several different higher-resolution assays in maize, a model species that is likely representative of many large crop genomes, revealed the importance of chromatin loops for influencing gene expression and phenotype (Li et al., 2019; Peng et al., 2019; Ricci et al., 2019; Sun et al., 2020). Collectively, these studies indicated that: (i) interactions between genes and proximal (<2 kb) and distal (>20 kb) ACRs (i.e., putative CREs) were common, and confirmed many genetically identified long-distance regulatory regions; (ii) genes with chromatin interactions associated with active promoters and enhancers tended to have higher expression levels than those without; (iii) functional CRE-gene interactions showed a strong loop signal intensity and tended to lie directly upstream of the gene (i.e., gene skipping was less common than direct contact; Ricci et al., 2019); (iv) gene pairs connected by loops within their proximal promoters were often transcriptionally coordinated; (v) tissue-specific (i.e., ear vs. shoot) proximal-distal interactions correlated with tissue-specific gene expression; and (vi) genes and CREs were often connected by multiple loops suggesting a complex pattern of regulation. Many of these features are likely to be conserved in other plant genomes and serve as a foundation for predicting functional regulatory elements in other species. However, given the vast diversity and size differences among plant genomes, and the prevalence of polyploidy among domesticated crop species, it is possible that many species exhibit unique chromatin conformation features that influence gene expression and certain species-specific traits (Wang et al., 2017; Concia et al., 2020).

Overall, these studies in plants confirm that long-range contacts do frequently occur in plants and raise many additional intriguing aspects of gene regulation. For example, chromatin contact mapping suggests that like in animals, gene expression can be influenced by multiple regulatory regions and that conversely, an individual regulatory region can modulate multiple genes (Wang et al., 2017; Ricci et al., 2019; Gasperini et al., 2020).

Understanding this complexity will likely shed light on prior genetic data and assist with future engineering efforts.

PROSPECTS FOR MINING REGULATORY DIVERSITY IN EXISTING GERMPLASM

De novo whole genome assembly is becoming wide available opening the door for mining regulatory diversity among not only many different plant species, but also closely related inbred lines, accessions, and varieties (Tao et al., 2019; Danilevicz et al., 2020). Such pan-genome collections allow for identification of regulatory variants including both coding and expression alleles including those associated with gene presence/absence, copy number variation, SNPs, indels, and structural variation, and are likely to be highly informative (Darracq et al., 2018; Sun et al., 2018; Gao et al., 2019; Yang et al., 2019a,b; Zhou et al., 2019; Alonge et al., 2020; Song et al., 2020). Similarly, understanding regulatory divergence among sub-genomes in polyploidy species is another exciting yet challenging prospect (Bao et al., 2019). Annotation of both conserved and accession-specific functional elements within these assemblies will likely require both empirical and machine learning based techniques (Michael and VanBuren, 2020). Among these annotation efforts, cataloging and characterizing CREs and individual TF binding events in plant genomes will be essential for understanding transcriptional and phenotypic variation. Much like the genetic maps and gene maps that have guided plant molecular genetics research for the past several decades, we envision that physical maps of annotated non-coding regulatory regions and CREs will be highly useful for both basic research and precision plant breeding. The generation of species-specific "genomic navigation systems" could transform research in much the same way that cellular navigation systems have enabled expanded and more efficient travel in everyday life. Ultimately, the ability to use CRISPR-based technologies to edit specific regulatory elements and alter transcriptional outputs offers great promise for engineering desirable traits (Rodríguez-Leal et al., 2017; Eshed and Lippman, 2019), providing new ways to increase genetic gain and affording a broader spectrum of genetic variation than what is seen in nature, ultimately transforming our approach to crop improvement.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

MG, FF, and AG are supported by a grant from the National Science Foundation (TRTech-PGR IOS1916804).

REFERENCES

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145.e23–161.e23. doi: 10.1016/j.cell.2020.05.021

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* 21, 71–87. doi: 10.1038/s41576-019-0173-8

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077. doi: 10.1126/science.1232542

Bao, Y., Hu, G., Grover, C. E., Conover, J., Yuan, D., and Wendel, J. F. (2019). Unraveling cis and trans regulatory evolution during cotton domestication. *Nat. Commun.* 10:5399. doi: 10.1038/s41467-019-13386-w

Bartlett, A., O'Malley, R. C., Huang, S. C., Galli, M., Nery, J. R., Gallavotti, A., et al. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* 12, 1659–1672. doi: 10.1038/nprot.2017.055

Betsiashvili, M., Ahern, K. R., and Jander, G. (2015). Additive effects of two quantitative trait loci that confer *Rhopalosiphum maidis* (corn leaf aphid) resistance in maize inbred line Mo17. *J. Exp. Bot.* 66, 571–578. doi: 10.1093/jxb/eru379

Bolduc, N., Yilmaz, A., Mejia-Guerra, M. K., Morohashi, K., O'Connor, D., Grotewold, E., et al. (2012). Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev.* 26, 1685–1690. doi: 10.1101/gad.193433.112

Brooks, M. D., Cirrone, J., Pasquino, A. V., Alvarez, J. M., Swift, J., Mittal, S., et al. (2019). Network walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. *Nat. Commun.* 10:1569. doi: 10.1038/s41467-019-09522-1

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24, 1242–1255. doi: 10.1105/tpc.111.095232

Castelletti, S., Tuberosa, R., Pindo, M., and Salvi, S. (2014). A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL vgt1. *G3 (Bethesda)* 4, 805–812. doi: 10.1534/g3.114.010686

Chen, K., Wang, Y., Zhang, R., Zhang, H., and Gao, C. (2019). CRISPR/Cas genome editing and precision plant breeding in agriculture. *Annu. Rev. Plant Biol.* 70, 667–697. doi: 10.1146/annurev-arplant-050718-100049

Concia, L., Veluchamy, A., Ramirez-Prado, J. S., Martin-Ramirez, A., Huang, Y., Perez, M., et al. (2020). Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biol.* 21:104. doi: 10.1186/s13059-020-01998-1

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., et al. (2020). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 23991–24000. doi: 10.1073/pnas.2010250117

Crisp, P. A., Noshay, J. M., Anderson, S. N., and Springer, N. M. (2019). Opportunities to use DNA methylation to distil functional elements in large crop genomes. *Mol. Plant* 12, 282–284. doi: 10.1016/j.molp.2019.02.006

Danilevicz, M. F., Tay Fernandez, C. G., Marsh, J. I., Bayer, P. E., and Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* 54, 18–25. doi: 10.1016/j.pbi.2019.12.005

Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J. P., Mary-Huard, T., et al. (2018). Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC Genomics* 19:119. doi: 10.1186/s12864-018-4490-7

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* 38, 56–65. doi: 10.1038/s41587-019-0315-8

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell* 166, 538–554. doi: 10.1016/j.cell.2016.07.012

Dong, Z., Alexander, M., and Chuck, G. (2019). Understanding grass domestication through maize mutants. *Trends Genet.* 35, 118–128. doi: 10.1016/j.tig.2018.10.007

Dong, P., Tu, X., Chu, P. Y., Lü, P., Zhu, N., Grierson, D., et al. (2017). 3D chromatin architecture of large plant genomes determined by local a/B compartments. *Mol. Plant* 10, 1497–1509. doi: 10.1016/j.molp.2017.11.005

Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C. T., Gent, J. I., Guo, L., et al. (2012). Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genet.* 8:e1003127. doi: 10.1371/journal.pgen.1003127

Eshed, Y., and Lippman, Z. B. (2019). Revolutions in agriculture chart a course for targeted breeding of old and new crops. *Science* 366:eaax0025. doi: 10.1126/science.aax0025

Feng, F., Qi, W., Lv, Y., Yan, S., Xu, L., Yang, W., et al. (2018). OPAQUE11 is a central hub of the regulatory network for maize endosperm development and nutrient metabolism. *Plant Cell* 30, 375–396. doi: 10.1105/tpc.17.00616

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462, 58–64. doi: 10.1038/nature08497

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., et al. (2018). The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.* 9:4526. doi: 10.1038/s41467-018-06977-6

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51, 1044–1051. doi: 10.1038/s41588-019-0410-2

Gasperini, M., Tome, J. M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* 21, 292–310. doi: 10.1038/s41576-019-0209-0

Gregis, V., Andrés, F., Sessa, A., Guerra, R. F., Simonini, S., Mateos, J. L., et al. (2013). Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. *Genome Biol.* 14:R56. doi: 10.1186/gb-2013-14-6-r56

Han, J. J., Jackson, D., and Martienssen, R. (2012). Pod corn is caused by rearrangement at the Tunicate1 locus. *Plant Cell* 24, 2733–2744. doi: 10.1105/tpc.112.100537

Hardison, R. C., and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* 13, 469–483. doi: 10.1038/nrg3242

Hernandez, J. M., Feller, A., Morohashi, K., Frame, K., and Grotewold, E. (2007). The basic helix-loop-helix domain of maize R links transcriptional regulation and histone modifications by recruitment of an EMSY-related factor. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17222–17227. doi: 10.1073/pnas.0705629104

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 157–165. doi: 10.1016/j.bbagrm.2016.05.010

Huang, S. -S. C., and Ecker, J. R. (2018). Piecing together cis-regulatory networks: insights from epigenomics studies in plants. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 10:e1411. doi: 10.1002/wsbm.1411

Huang, W., Pérez-García, P., Pokhilko, A., Millar, A. J., Antoshechkin, I., Riechmann, J. L., et al. (2012). Mapping the core of the *Arabidopsis* circadian clock defines the network structure of the oscillator. *Science* 336, 75–79. doi: 10.1126/science.1219075

Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X., et al. (2017). ZmCCT9 enhances maize adaptation to higher latitudes. *Proc. Natl. Acad. Sci. U. S. A.* 115, E334–E341. doi: 10.1073/pnas.1718058115

Jo, L., Pelletier, J. M., Hsu, S. W., Baden, R., Goldberg, R. B., and Harada, J. J. (2020). Combinatorial interactions of the LEC1 transcription factor specify diverse developmental programs during soybean seed development. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1223–1232. doi: 10.1073/pnas.1918441117

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo-DNA interactions. *Science* 316, 1497–1502. doi: 10.1126/science.1141319

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-binding specificities of human transcription factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009

Jores, T., Tonnies, J., Dorrity, M. W., Cuperus, J., Fields, S., and Queitsch, C. (2020). Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell* 32, 2120–2131. doi: 10.1105/tpc.20.00155

Jung, J. -H., Domijan, M., Klose, C., Biswas, S., Ezer, D., Gao, M., et al. (2016). Phytochromes function as thermosensors in *Arabidopsis*. *Science* 354, 886–889. doi: 10.1126/science.aaf6005

Kaufmann, K., Muiño, J. M., Østerås, M., Farinelli, L., Krajewski, P., and Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nat. Protoc.* 5, 457–472. doi: 10.1038/nprot.2009.244

Kidder, B. L., Hu, G., and Zhao, K. (2011). ChIP-seq: technical considerations for obtaining high-quality data. *Nat. Immunol.* 12, 918–922. doi: 10.1038/ni.2117

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 20, 207–220. doi: 10.1038/s41576-018-0089-8

Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science* 304:982. doi: 10.1126/science.1095011

Kong, Q., Pattanaik, S., Feller, A., Werkman, J. R., Chai, C., Wang, Y., et al. (2012). Regulatory switch enforced by basic helix-loop-helix and ACT-domain mediated dimerizations of the maize transcription factor R. *Proc. Natl. Acad. Sci.* 109, E2091–E2097. doi: 10.1073/pnas.1205513109

Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., et al. (2006). An SNP caused loss of seed shattering during rice domestication. *Science* 312, 1392–1396. doi: 10.1126/science.1126410

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Lau, O. S., Davies, K. A., Chang, J., Adrian, J., Rowe, M. H., Ballenger, C. E., et al. (2014). Direct roles of SPEECHLESS in the specification of stomatal self-renewing cells. *Science* 345, 1605–1609. doi: 10.1126/science.1256888

Lee, J. H., Jin, S., Kim, S. Y., Kim, W., and Ahn, J. H. (2017). A fast, efficient chromatin immunoprecipitation method for studying protein-DNA binding in *Arabidopsis* mesophyll protoplasts. *Plant Methods* 13:42. doi: 10.1186/s13007-017-0192-4

Li, E., Liu, H., Huang, L., Zhang, X., Dong, X., Song, W., et al. (2019). Long-range interactions between proximal and distal regulatory regions in maize. *Nat. Commun.* 10:2633. doi: 10.1038/s41467-019-10603-4

Li, C., Qiao, Z., Qi, W., Wang, Q., Yuan, Y., Yang, X., et al. (2015). Genome-wide characterization of *cis*-acting DNA targets reveals the transcriptional regulatory framework of *Opaque2* in maize. *Plant Cell* 27, 532–545. doi: 10.1105/tpc.114.134858

Liu, C., Cheng, Y. J., Wang, J. W., and Weigel, D. (2017). Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat. Plants* 3, 742–748. doi: 10.1038/s41477-017-0005-9

Liu, W., and Stewart, C. N. (2016). Plant synthetic promoters and transcription factors. *Curr. Opin. Biotechnol.* 37, 36–44. doi: 10.1016/j.copbio.2015.10.001

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., and Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* 5, 1250–1259. doi: 10.1038/s41477-019-0548-z

Lü, P., Yu, S., Zhu, N., Chen, Y. R., Zhou, B., Pan, Y., et al. (2018). Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat. Plants* 4, 784–791. doi: 10.1038/s41477-018-0249-z

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., et al. (2018). Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell* 30, 15–36. doi: 10.1105/tpc.17.00581

Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., et al. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* 38, 948–952. doi: 10.1038/ng1841

Marand, A. P., Zhang, T., Zhu, B., and Jiang, J. (2017). Towards genome-wide prediction and characterization of enhancers in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 131–139. doi: 10.1016/j.bbagrm.2016.06.006

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., et al. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461, 1135–1138. doi: 10.1038/nature08498

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi: 10.1038/nrg3605

Michael, T. P., and VanBuren, R. (2020). Building near-complete plant genomes. *Curr. Opin. Plant Biol.* 54, 26–33. doi: 10.1016/j.pbi.2019.12.009

Mumbach, M., Rubin, A., Flynn, R., Dai, C., Khavari, P., Greenleaf, W., et al. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* 13, 919–922. doi: 10.1038/nmeth.3999

Noshay, J. M., Marand, A. P., Anderson, S. N., Zhou, P., Guerra, M. K. M., Lu, Z., et al. (2020). Cis-regulatory elements within TEs can influence expression of nearby maize genes. bioRxiv [Preprint]. doi: 10.1101/2020.05.20.107169

O'Malley, R. C., Huang, S. S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., et al. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 165, 1280–1292. doi: 10.1016/j.cell.2016.04.038

Oka, R., Zicola, J., Weber, B., Anderson, S. N., Hodgman, C., Gent, J. I., et al. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* 18:137. doi: 10.1186/s13059-017-1273-4

Olsen, K. M., and Wendel, J. F. (2013). Crop plants as models for understanding plant adaptation and diversification. *Front. Plant Sci.* 4:290. doi: 10.3389/fpls.2013.00290

Para, A., Li, Y., Marshall-Colón, A., Varala, K., Francoeur, N. J., Moran, T. M., et al. (2014). Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 111, 10371–10376. doi: 10.1073/pnas.1404657111

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi: 10.1038/nrg2641

Parvathaneni, R. K., Bertolini, E., Shamimuzzaman, M., Vera, D. L., Lung, P. -Y., Rice, B. R., et al. (2020). The regulatory landscape of early maize inflorescence development. *Genome Biol.* 21:165. doi: 10.1186/s13059-020-02070-8

Pautler, M., Eveland, A. L., Larue, T., Yang, F., Weeks, R., Lunde, C., et al. (2015). FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. *Plant Cell* 27, 104–120. doi: 10.1105/tpc.114.132506

Peng, Y., Xiong, D., Zhao, L., Ouyang, W., Wang, S., Sun, J., et al. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat. Commun.* 10:2632. doi: 10.1038/s41467-019-10602-5

Pliner, H. A., Packer, J. S., Steemers, F. J., Shendure, J., Aghamirzaie, D., Srivatsan, S., et al. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data technology. *Mol. Cell* 71, 858.e8–871.e8. doi: 10.1016/j.molcel.2018.06.044

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nat. Plants* 5, 1237–1249. doi: 10.1038/s41477-019-0547-0

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., and Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci.* 113, E3177–E3184. doi: 10.1073/pnas.1525244113

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., and Lippman, Z. B. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell* 171, 470.e8–480.e8. doi: 10.1016/j.cell.2017.08.030

Rossi, M., Duffy, T., Conti, G., Almeida, J., Bermudez, L., Fernie, A. R., et al. (2014). Natural occurring epialleles determine vitamin E accumulation in tomato fruits. *Nat. Commun.* 5:3027. doi: 10.1038/ncomms5027

Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., et al. (2017). Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* 67, 837.e7–852.e7. doi: 10.1016/j.molcel.2017.07.022

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286. doi: 10.1038/nrg3682

Skene, P. J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6:e21856. doi: 10.7554/eLife.21856

Skene, P. J., Henikoff, J. G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* 13, 1006–1019. doi: 10.1038/nprot.2018.015

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* 39, 381–399. doi: 10.1016/j.tibs.2014.07.002

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* 6, 34–45. doi: 10.1038/s41477-019-0577-7

Song, L., Huang, S. -S. C., Wise, A., Castanon, R., Nery, J. R., Chen, H., et al. (2016). A transcription factor hierarchy defines an environmental stress response network. *Science* 354:aag1550. doi: 10.1126/science.aag1550

Spitz, F., and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626. doi: 10.1038/nrg3207

Springer, N., de León, N., and Grotewold, E. (2019). Challenges of translating gene regulatory information into agronomic improvements. *Trends Plant Sci.* 24, 1075–1082. doi: 10.1016/j.tplants.2019.07.004

Stam, M., Belele, C., Dorweiler, J. E., and Chandler, V. L. (2002). Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes Dev.* 16, 1906–1918. doi: 10.1101/gad.1006702

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* 43, 1160–1163. doi: 10.1038/ng.942

Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., et al. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* 8, 2015–2030. doi: 10.1016/j.celrep.2014.08.019

Sun, Y., Dong, L., Zhang, Y., Lin, D., Xu, W., Ke, C., et al. (2020). 3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize. *Genome Biol.* 21:143. doi: 10.1186/s13059-020-02063-7

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., et al. (2018). Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* 50, 1289–1295. doi: 10.1038/s41588-018-0182-0

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., et al. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 24, 1963–1976. doi: 10.1101/gr.168872.113

Swinnen, G., Goossens, A., and Pauwels, L. (2016). Lessons from domestication: targeting cis-regulatory elements for crop improvement. *Trends Plant Sci.* 21, 506–515. doi: 10.1016/j.tplants.2016.01.014

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant* 12, 156–169. doi: 10.1016/j.molp.2018.12.016

Todeschini, A. L., Georges, A., and Veitia, R. A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet.* 30, 211–219. doi: 10.1016/j.tig.2014.04.002

Tsuda, K., Kurata, N., Ohyanagi, H., and Hake, S. (2014). Genome-wide study of *KNOX* regulatory network reveals brassinosteroid catabolic genes important for shoot meristem function in rice. *Plant Cell* 26, 3488–3500. doi: 10.1105/tpc.114.129122

Tu, X., Mejía-Guerra, M. K., Franco, J. A. V., Tzeng, D., Chu, P.-Y., Dai, X., et al. (2020). The transcription regulatory code of a plant leaf. *Nat. Commun.* 11:5089. doi: 10.1038/s41467-020-18832-8

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* 28, 1089–1095. doi: 10.1038/nbt.1680

Wang, S., Li, S., Liu, Q., Wu, K., Zhang, J., Wang, S., et al. (2015b). The OsSPL16-GW7 regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* 47, 949–954. doi: 10.1038/ng.3352

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., et al. (2015a). Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res.* 25, 246–256. doi: 10.1101/gr.170332.113

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807

Wang, X., Wang, H., Liu, S., Ferjani, A., Li, J., Yan, J., et al. (2016). Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. *Nat. Genet.* 48, 1233–1241. doi: 10.1038/ng.3636

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., et al. (2019b). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* 10:2989. doi: 10.1038/s41467-019-10820-x

Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., et al. (2013). CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16969–16974. doi: 10.1073/pnas.1310949110

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., et al. (2019a). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* 51, 1052–1059. doi: 10.1038/s41588-019-0427-6

Yoo, S. D., Cho, Y. H., and Sheen, J. (2007). *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* 2, 1565–1572. doi: 10.1038/nprot.2007.199

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A. P., Wu, Y., et al. (2018). Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiol.* 176, 2789–2803. doi: 10.1104/pp.17.01467

Zheng, L., McMullen, M. D., Bauer, E., Schön, C. C., Gierl, A., and Frey, M. (2015). Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in *Zea mays*. *J. Exp. Bot.* 66, 3917–3930. doi: 10.1093/jxb/erv192

Zhong, S., Fei, Z., Chen, Y. R., Zheng, Y., Huang, M., Vrebalov, J., et al. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat. Biotechnol.* 31, 154–159. doi: 10.1038/nbt.2462

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., et al. (2019). The population genetics of structural variants in grapevine domestication. *Nat. Plants* 5, 965–979. doi: 10.1038/s41477-019-0507-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 November 2020
doi: 10.3389/fgene.2020.577897

Alternative Splicing Diversified the Heat Response and Evolutionary Strategy of Conserved *Heat Shock Protein 90s* in Hexaploid Wheat (*Triticum aestivum* L.)

Yunze Lu[1,2], *Peng Zhao*[1], *Aihua Zhang*[1], *Lingjian Ma*[1], *Shengbao Xu*[1]* and *Xiaoming Wang*[1]*

[1] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Agronomy, Northwest A&F University, Yangling, China, [2] School of Landscape and Ecological Engineering, Hebei University of Engineering, Handan, China

OPEN ACCESS

Edited by:
Yin Li,
Rutgers, The State University of New Jersey, United States

Reviewed by:
Nelson Garcia,
University of Minnesota Twin Cities, United States
Zhaorong Hu,
China Agricultural University, China
Lingli Dong,
Institute of Genetics and Developmental Biology (CAS), China

****Correspondence:***
Shengbao Xu
xushb@nwsuaf.edu.cn
Xiaoming Wang
wangxm@nwsuaf.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 June 2020*
Accepted: *29 October 2020*
Published: *27 November 2020*

Citation:
Lu Y, Zhao P, Zhang A, Ma L, Xu S and Wang X (2020) Alternative Splicing Diversified the Heat Response and Evolutionary Strategy of Conserved Heat Shock Protein 90s in Hexaploid Wheat (Triticum aestivum L.). Front. Genet. 11:577897. doi: 10.3389/fgene.2020.577897

Crops are challenged by the increasing high temperature. Heat shock protein 90 (HSP90), a molecular chaperone, plays a critical role in the heat response in plants. However, the evolutionary conservation and divergence of *HSP90s* homeologs in polyploidy crops are largely unknown. Using the newly released hexaploid wheat reference sequence, we identified 18 *TaHSP90s* that are evenly distributed as homeologous genes among three wheat subgenomes, and were highly conserved in terms of sequence identity and gene structure among homeologs. Intensive time-course transcriptomes showed uniform expression and transcriptional response profiles among the three *TaHSP90* homeologs. Based on the comprehensive isoforms generated by combining full-length single-molecule sequencing and Illumina short read sequencing, 126 isoforms, including 90 newly identified isoforms of *TaHSP90s*, were identified, and each *TaHSP90* generated one to three major isoforms. Intriguingly, the numbers and the splicing modes of the major isoforms generated by three *TaHSP90* homeologs were obviously different. Furthermore, the quantified expression profiles of the major isoforms generated by three *TaHSP90* homeologs are also distinctly varied, exhibiting differential alternative splicing (AS) responses of homeologs. Our results showed that the AS diversified the heat response of the conserved *TaHSP90s* and provided a new perspective for understanding about functional conservation and divergence of homologous genes.

Keywords: wheat, heat stress, HSP90, alternative splicing, evolutionary conservation and divergence

INTRODUCTION

As a result of global warming and increasing frequent short episodes of extreme high temperature (Liu et al., 2014; Prasad and Djanaguiraman, 2014), heat stress has become one of the major factors that limit crop production and quality (Asseng et al., 2015; Lesk et al., 2016; Zhao et al., 2017). Each degree Celsius increase in global mean temperature would reduce crop yield by 6–10% under heat stress (Lesk et al., 2016; Zhao et al., 2017).

Plants have evolved complex systems to cope with heat stress. Heat shock proteins (HSPs), which function as molecular chaperones, have well-established roles in the heat response (Wang et al., 2004). Ninety kDa HSPs (HSP90s) function not only in protein folding, degradation, and transportation, as characteristic of chaperones, but also in signaling transduction and protein kinase activity regulation, in an ATP-dependent manner (Carretero-Paulet et al., 2013; Schopf et al., 2017). Under heat stress, the 70-kDa HSP/HSP90 complex disassociates and releases the master heat response regulator, heat shock transcription factor A1s (HSFA1s), resulting in the activation of the plant heat-responsive transcription cascade (Hahn et al., 2011; Ohama et al., 2017). More recently, HSP90s have been proven to have the ability to stabilize the circadian clock ZEITLUPE and auxin cofactor F-box protein to maintain plant growth under heat stress (Wang et al., 2016; Gil et al., 2017). Under normal conditions, the downregulated expression of Arabidopsis HSP90s results in abnormal growth and development, such as embryo defects (Queitsch et al., 2002). HSP90s are also considered to exert capacitor-buffering effects of genetic perturbation, such as genetic variations and mutations, on plant morphology and phenotype (Rutherford and Lindquist, 1998; Geiler-Samerotte et al., 2016).

The identification and expression pattern analysis of *HSP90*s have been widely reported in Arabidopsis, rice, chickpea, and pigeonpea (Krishna and Gloor, 2001; Swindell et al., 2007; Hu et al., 2009; Agarwal et al., 2016). In Arabidopsis, seven *AtHSP90s* were characterized and classified into different subfamilies based on their subcellular localization. Under normal condition, *HSP90AA* (*AtHSP90-1*) and *HSP90ABs* (*AtHSP90-2-AtHSP90-4*) are located in the cytoplasm, and *HSP90C1* (*AtHSP90-5*), *HSP90C2* (*AtHSP90-6*), and *HSP90B* (*AtHSP90-7*) are located in chloroplasts, mitochondria, and endoplasmic reticulum, respectively. Under heat stress, HSP90s were found to translocate into the nucleus, to regulate the expression of heat stress-responsive genes (Meiri and Breiman, 2009; Wang et al., 2016). The *HSP90AAs* are highly heat inducible, whereas *HSP90ABs* are constitutively expressed in Arabidopsis and rice (Yabe et al., 1994; Swindell et al., 2007; Hu et al., 2009). Similar heat response patterns of *HSP90s* within a subfamily are observed in *Populus* and allotetraploid tobacco (Zhang et al., 2013; Song et al., 2019). Meanwhile, previous studies also demonstrate the conservation in terms of exon–intron structures and protein motifs of *HSP90s* within a subfamily during plant evolution (Chen et al., 2006; Xu et al., 2012; Zhang et al., 2017).

Hexaploid wheat (*Triticum aestivum* L.), a major crop worldwide, contains three subgenomes, AA, BB, DD. These three subgenomes derive from three diploids through two major hybridization events. The first hybridization occurs between the diploid *T. urartu* (AA progenitor) and an unknown diploid Aegilops species (BB progenitor), possibly *Aegilops sharonensis* or *Ae. speltoides*, leading to the emergence of allotetraploid wild emmer wheat (*T. turgidum* ssp. *Dicoccoides*, AABB). The wild emmer wheat subsequently evolves into durum wheat (*T. turgidum* ssp. *Durum*, AABB). The second hybridization between tetraploid emmer wheat and the diploid *Ae. tauschii* (DD progenitor) results in the hexaploid wheat (AABBDD) (Dubcovsky and Dvorak, 2007; Marcussen et al., 2014). Due to the short polyploidy history, the hexaploid wheat is used as a model in the field of polyploidy evolution study, and the functional conservation and divergence of the homeologous genes in hexaploid wheat have always been intriguing (Borrill et al., 2015; Ramírez-González et al., 2018). At transcriptional level, expression partitioning of wheat homeologous genes is commonly observed, and the ratio of partially expressed genes ranges from about 55% in normal condition to around 68% in stress condition such as heat, drought, and salt (Leach et al., 2014; Liu et al., 2015; Zhang et al., 2016). Thus, expression partitioning of homeologous genes is thought to be a common strategy for abiotic stress acclimation in hexaploid wheat (Liu et al., 2015; Zhang et al., 2016).

In hexaploid wheat, nine cytosolic *TaHSP90*s have been reported, of which three *TaHSP90AA*s are highly expressed in the reproductive organs that are necessary for seedling growth, while six *TaHSP90AB*s are constitutively expressed that are essential for disease resistance (Wang et al., 2011). However, the evolution process of *HSP90s* during wheat polyploidization and the functional conversation and divergence between *TaHSP90* homeologs in hexaploid wheat remain largely limited.

With posttranscriptional regulation, many heat response genes, including several types of transcription factors and *HSPs*, change their alternative splicing (AS) patterns and generate new isoforms, expanding the diversity of proteome and regulation modes in the heat response (Jiang et al., 2017; Keller et al., 2017; Liu et al., 2018). For example, under normal conditions, *HSFA2* mainly encodes a truncated isoform without transcriptional activation activity, while under heat stress, the intact and transcriptionally active isoform is largely expressed to induce the expression of heat response genes (Sugio et al., 2009; Cheng et al., 2015). Interestingly, a new, small truncated form is induced, and this results in the autoregulation of *HSFA2* under severe heat stress (Liu et al., 2013). Whether and how *HSP90*s respond to heat stress with posttranscriptional regulation remains to be determined. If it is, what posttranscriptional regulation means for *HSP90s* is intriguing. The advantages of full-length single-molecule sequencing with long sequencing read provide powerful tool for accurate isoform detection and could be suitable to answer this question.

In this study, we performed a genome-wide identification of the *TaHSP90s* in hexaploid wheat using the newly released wheat reference sequence (IWGSC RefSeq v1.0) (International Wheat Genome Sequencing Consortium [IWGSC], 2018) and investigated the transcriptional and AS reprogramming of *TaHSP90s* using the dynamic and intensive heat response transcriptomes generated by combining full-length single-molecule sequencing and Illumina short read sequencing in our previous study (Wang et al., 2019). The 18 *TaHSP90s* were highly conserved in terms of sequence and transcriptional response pattern among *TaHSP90* homeologs, while the number, the splicing modes, and the AS responses of the major isoforms generated by *TaHSP90s* homeologs were distinctly different. Our findings indicated that AS regulation diversified the heat response of the conserved *TaHSP90* homeologs and possibly facilitated the evolutionary divergence of *TaHSP90* homeologs.

MATERIALS AND METHODS

Identification of HSP90s in Hexaploid Wheat and Its Progenitors

Genome and protein sequences and genome annotation files of hexaploid wheat were downloaded from URGI[1], and the protein sequences of *Ae. tauschii*, *T. urartu*, durum wheat, and wild emmer wheat were downloaded from the EnsemblPlants database[2]. Genome and protein sequences of *Ae. sharonensis* and *Ae. speltoides* were obtained from PGSB PlantsDB[3]. A blastp search was performed against the protein sequences of the above species using the HSP90 sequences of Arabidopsis and rice as queries with the following parameters: an e-value lower than 1e-5 and an identity score above 50%. Using the HMM profiles of the HSP90 domain (PF00183) downloaded from the Pfam database, the Hmmsearch engine in the HMMER3.0 program was also used to search these proteins with a threshold of 1e-5. Then, the blastp and HMMER results were merged, and redundancy was removed. All of the obtained HSP90 candidate sequences were subjected to search against the SMART database[4] (Letunic and Bork, 2018) to manually confirm the presence of the histidine kinase-like ATPases (HATPase_c) and HSP90 domains (Zhang et al., 2013; Agarwal et al., 2016). Candidates that contained both the HATPase_c and HSP90 domain were regarded as HSP90s.

Phylogenetic Relationship Analysis

Multiple sequence alignment was performed by MAFFT (L-INS-I algorithm) program using the protein sequences. Sequences from Arabidopsis and rice were used as markers to clarify the phylogeny. To investigate the evolutionary relationship of HSP90, a maximum likelihood tree was constructed using IQ-TREE (Nguyen et al., 2015). The substitution model was calculated with ModelFinder (integrated in IQ-TREE; best-fit model: JTT + G4 chosen according to the Bayesian information criterion) (Kalyaanamoorthy et al., 2017). The phylogenetic tree was examined by Ultrafast bootstraps (with parameters "-bb 1,000 -bnni") as well as a Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT, with parameters "-alrt 1,000") (Guindon et al., 2010; Minh et al., 2013; Hoang et al., 2018). The tree file was visualized by Interactive Tree Of Life (iTOL) v4[5] (Letunic and Bork, 2019).

Naming of Hexaploid Wheat *HSP90s*

For clarity, we renamed all the *HSP90s* of hexaploid wheat, taking into account the naming convention of *HSP90* family members in Arabidopsis and rice. Each gene name started with the abbreviation of the hexaploid wheat (*T. aestivum*, *Ta*), followed by the abbreviation of this gene family (*HSP90*) and the subfamily name that is in Arabidopsis (AA, AB, B, C1, and C2), and finally ended with the chromosome and subgenome information. For example, the name of gene with geneID "*TraesCS2A01G033700*" was renamed as *TaHSP90AA-2A*; in other words, it was a *HSP90* gene belonging to the subfamily AA, and it was located on the AA subgenome on chromosome group 2. *HSP90s* from progenitors of hexaploid wheat were only presented in the phylogenetic relationship analysis; thus, they were not renamed.

Transcriptional Regulation and AS Analysis of *TaHSP90s*

The transcriptome data were obtained from our previous research (Wang et al., 2019). Briefly, hexaploid wheat plants (*T. aestivum* cv. Chinese Spring) were grown in greenhouse under normal condition. After 15 days from anthesis, the plants were subjected to heat stress (37/17°C, 14/10 h). The filling grains and flag leaves were sampled at different time points and then for RNA isolation. The RNA-seq data, generated by Illumina (150 bp paired-end sequencing by Illumina's HiSeq X Ten platform) and PacBio (PacBio RS II platform) sequencing, and data processing, isoform characterization were performed as described in a previous study (Wang et al., 2019). Briefly, the PacBio sequencing full-length and non-chimeric (FLNC) reads were first mapped to the bread wheat reference genome, and the mapped FLNC reads were filtered and corrected with our previously defined criterion. Then, the FLNC reads that were mapped to the same genome loci and shared the same splicing junctions were collapsed into one isoform. Finally, the identified isoforms were further filtered based on the number of FLNC reads supporting this isoform, percentage-of-identity of FLNC read-genome alignments, and the junction site verifications, whether they were supported by Illumina reads or genomic annotations as our previous description. The abundances of genes and isoforms were calculated by fragments per kilobase of exon per million fragment mapped (FPKM) values. The heat maps were drawn by "pheatmap" package in R software (version 3.6.1) with log2-transformed (FPKM + 1) values (Kolde and Kolde, 2015). Differentially expressed genes were identified by EdgeR (Robinson et al., 2010), and the genes that displayed fold change ≥ 2 and an FDR adjusted P-value < 0.01 were defined as differentially expressed genes, with non-stressed samples as the control. The gene with either of the following situation was regarded as a differentially spliced gene: (i) isoform set (FPKM ≥ 1 and fully supported by Illumina reads) differed between heat stress and control sample. (ii) The isoform expression percentage (IEP) changed by more than 30% between the control and heat stress samples (Wang et al., 2019).

RESULTS

HSP90 Genes in Hexaploid Wheat and Its Progenitors

As a result of sequence searches and domain confirmation, 5, 5, 4, 5, 11, 13, and 18 *HSP90s* were identified in *T. urartu* (AA progenitor), *Ae. sharonensis* (possible BB progenitor), *Ae. speltoides* (possible BB progenitor), *Ae. tauschii* (DD progenitor), wild emmer wheat, durum wheat, and hexaploid

[1] https://urgi.versailles.inra.fr/

[2] http://plants.ensembl.org/

[3] ftp://ftpmips.helmholtz-muenchen.de/plants/

[4] http://smart.embl-heidelberg.de/

[5] https://itol.embl.de/

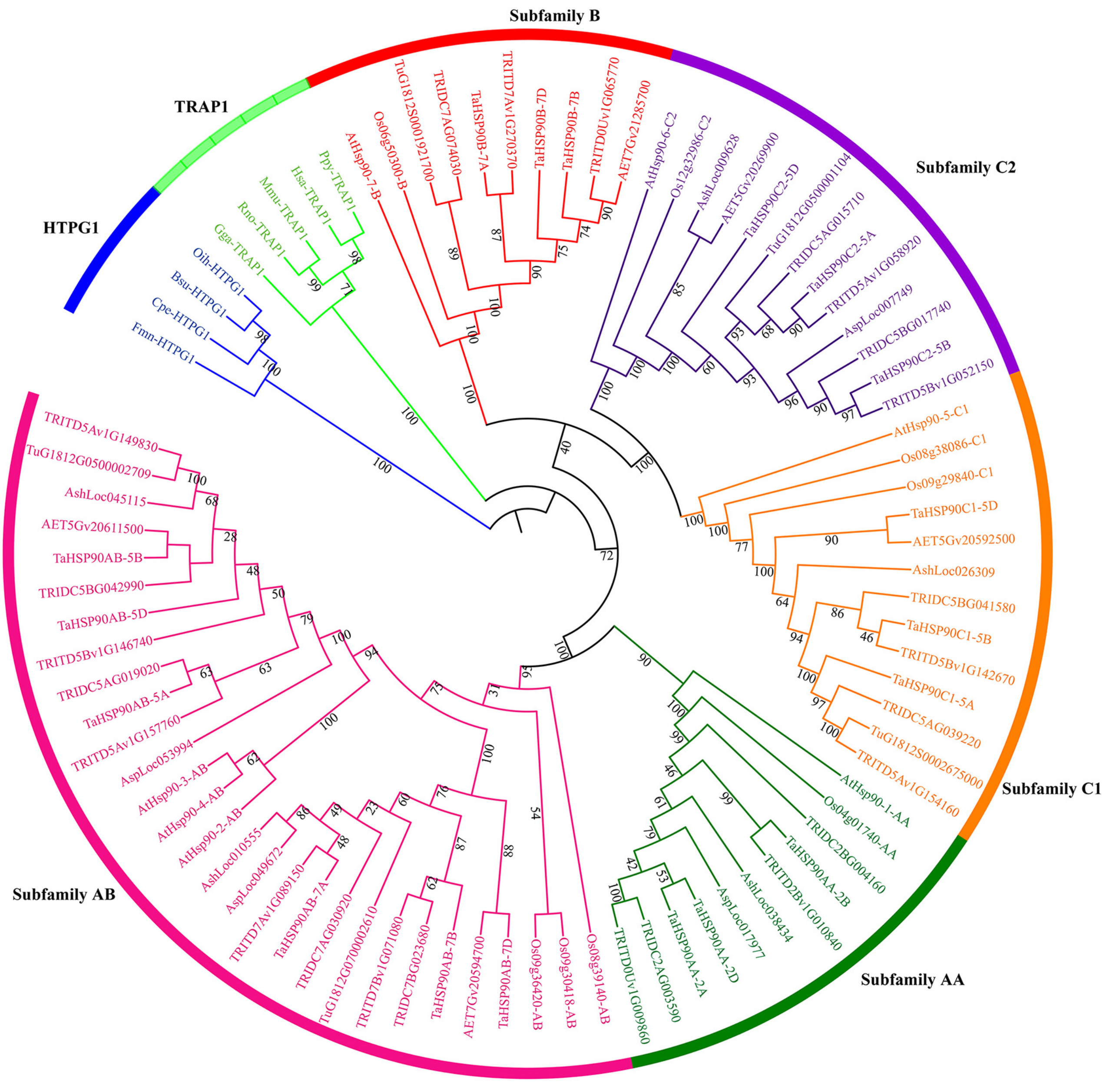


FIGURE 1 | Maximum likelihood phylogeny of HSP90 proteins. HSP90 sequences were from Arabidopsis, rice, *T. urartu* (Tu), *Ae. sharonensis* (Ash), *Ae. speltoides* (Asp), *Ae. tauschii* (Aet), wild emmer wheat (*T. turgidum* ssp. *dicoccoides*, TRIDC), durum wheat (*T. turgidum* ssp. *durum*, TRITD), and hexaploid wheat (*T. aestivum*, Ta). The phylogenetic tree was inferred by IQ-TREE with Ultrafast bootstraps (1,000 replicates) and visualized by the Interactive Tree Of Life (Itol). Subfamilies were classified according to Arabidopsis and rice HSP90 subfamilies and showed by different colors. Bootstrap values were shown near the branches. HTPG1 (high-temperature protein G 1), the HSP90 homeolog in Eubacteria, and TRAP1 (tumor necrosis factor receptor-associated protein 1), which was the mitochondrial HSP90 homeolog in Animalia, were set as outgroups according to Chen et al. (2006). Bsu, *Bacillus subtilis*; Cpe, *Clostridium perfringens*; Fmn, *Fusobacterium nucleatum*; Gga, *Gallus gallus*; Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Oih, *Oceanobacillus iheyensis*; Ppy, *Pongo pygmaeus*; Rno, *Rattus norvegicus*.

wheat genome, respectively. According to the phylogenetic tree, these *HSP90s* were clearly classified into subfamily AA, AB, B, C1, and C2 referred to as the classification of Arabidopsis and rice (**Figure 1**).

The clear polyploidization process of hexaploid wheat provided an opportunity to study the evolution of *HSP90s* during this process (**Figure 2A**) (Dubcovsky and Dvorak, 2007; Marcussen et al., 2014). The numbers of *HSP90s* were around five for the diploid species, and the number in allotetraploid wild emmer wheat (11) seemed to be the sum of the two diploid progenitors. However, the number of *HSP90s* in hexaploid wheat (18) was somehow not the sum of the allotetraploid wild emmer wheat (11) and the diploid *Ae. tauschii* (5) (**Figure 2**). For example, the hexaploid and the allotetraploid wheat contain three and two *HSP90AA* members, respectively, but the *HSP90AA* members were absent in diploid progenitor

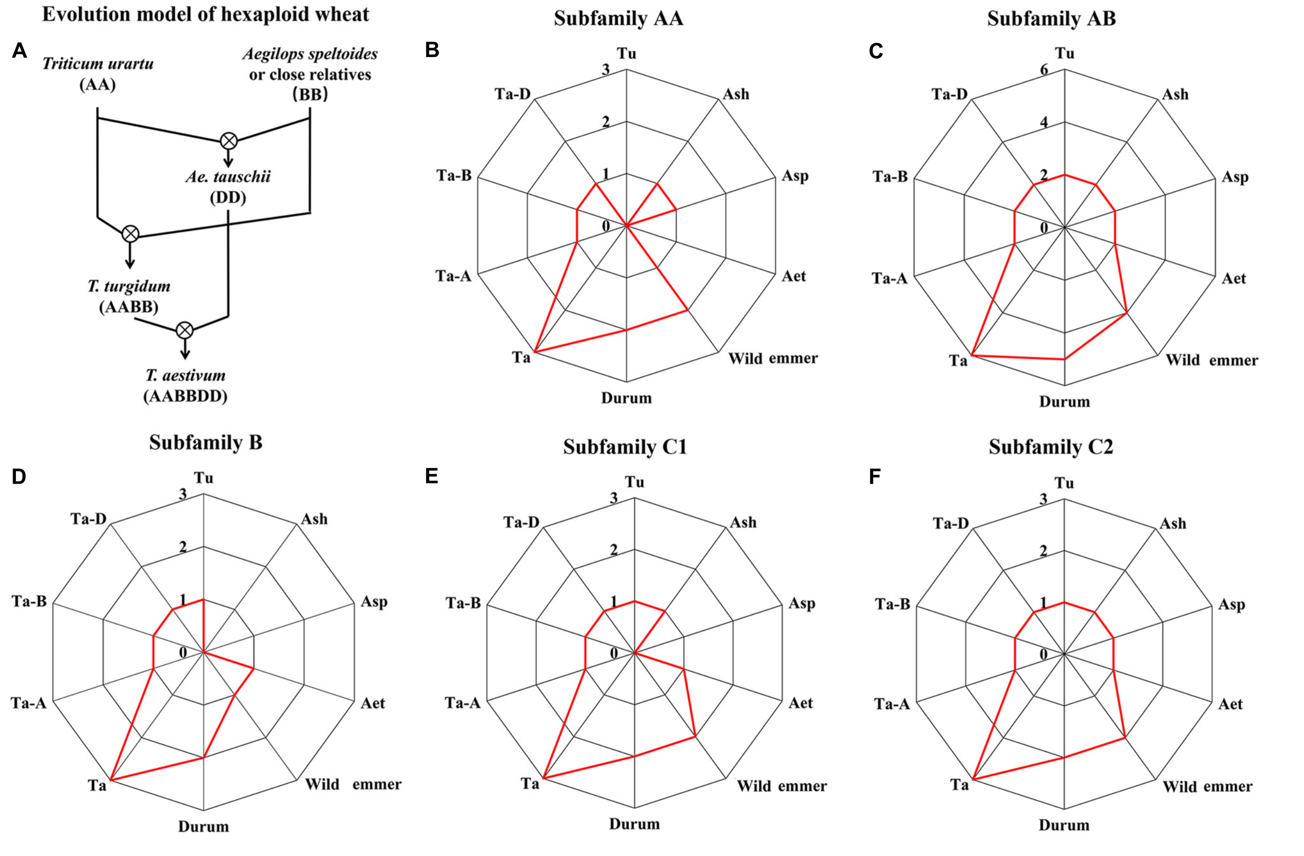


FIGURE 2 | Copy number of each HSP90 subfamily in seven Triticeae species. **(A)** Evolution model of hexaploid wheat. The symbol "⊗" indicated a hybridization event. The copy number of HSP90s in different species is shown as red lines in **(B)** for subfamily AA, **(C)** for subfamily AB, **(D)** for subfamily B, **(E)** for subfamily C1, and **(F)** for subfamily C2. Abbreviations: Tu, *T. urartu*; Ash, *Ae. sharonensis*; Asp, *Ae. speltoides*; Aet, *Ae. tauschii*; Wild emmer, *T. turgidum* ssp. *dicoccoides*; Durum, *T. turgidum ssp. durum*; Ta, *T. aestivum*. Ta-A, -B, -D indicated subgenome AA, BB, and DD of hexaploid wheat, respectively.

T. urartu and *Ae. tauschii* genomes. Syntenic analysis showed that the genomic segment containing the *HSP90AA* members and other three neighborhood genes were absent in the two diploid genomes (**Supplementary Figure 1**), suggesting hexaploid, and the allotetraploid wheat may acquire the gene copy during polyploidization, or the diploid genomes were not completely assembled. Conclusively, the copy number variation of *HSP90* was not completely consistent with the polyploidy level.

TaHSP90s Are Highly Conserved Between Three Homeologs

The 18 identified *HSP90s* in hexaploid wheat (TaHSP90s) were present in homeologous chromosome groups 2, 5, and 7 and were classified into *TaHSP90AAs* (3), *TaHSP90ABs* (6), *TaHSP90Bs* (3), *TaHSP90C1s* (3), and *TaHSP90C2s* (3) based on the phylogenetic tree (**Figure 1** and **Table 1**). Intriguingly, the *TaHSP90s* in each subfamily were evenly distributed among the AA, BB, and DD subgenomes, and the three *TaHSP90s* on the same chromosome group in each subfamily were regarded as the three *TaHSP90* homeologs here after.

Sequence analysis showed that the protein sequence identities of the three *TaHSP90* homeologs were above 96% (**Supplementary Table 1**). Accordingly, the protein sequence motifs (**Supplementary Figure 2A** and **Supplementary Table 2**) and gene structures (**Supplementary Figure 2B**) were also highly consistent among three TaHSP90 homeologs. Particularly, all the three *TaHSP90* homeologs contained the same exon number, and that is 3, 3, 15, 19, and 20 for subfamily AA, AB, B, C1, and C2, respectively. These results showed that the sequences and gene structures are highly conserved within three *TaHSP90s* homeologs in hexaploid wheat, which were consistent with the reports in *Populus* and *Brachypodium distachyon* (Zhang et al., 2013, 2017).

Conserved Heat Response Pattern Among Three *TaHSP90* Homeologs

Using the dynamic and intensive heat response transcriptomes of filling grains and flag leaves of wheat generated by our previous study (Wang et al., 2019), we investigated the heat response patterns of *TaHSP90s* and found that the heat stress response trend of each member of three *TaHSP90* homeologs was largely conserved, although the expression abundance was slightly different (**Figure 3A** and **Supplementary Figure 3**). For example, all the three *TaHSP90AAs* homeologs were lowly expressed under

TABLE 1 | Summary information of *TaHSP90s*.

TaHSP90s	GeneID	Exon number	Isoform number	FPKM values	
				Grains	Flag leaves
TaHSP90AA-2A	*TraesCS2A01G033700*	3	5	22–363	2–1,896
TaHSP90AA-2B	*TraesCS2B01G047400*	3	6	9–275	2–1,486
TaHSP90AA-2D	*TraesCS2D01G033200*	3	8	50–541	18–2,496
TaHSP90AB-7A	*TraesCS7A01G242200*	3	10	61–445	148–2,249
TaHSP90AB-7B	*TraesCS7B01G149200*	3	13	163–555	160–2,045
TaHSP90AB-7D	*TraesCS7D01G241100*	3	6	83–378	94–1,361
TaHSP90AB-5A	*TraesCS5A01G260600*	3	6	159–241	141–350
TaHSP90AB-5B	*TraesCS5B01G258900*	3	9	79–181	88–589
TaHSP90AB-5D	*TraesCS5D01G268000*	3	7	91–222	86–632
TaHSP90B-7A	*TraesCS7A01G529900*	15	7	19–59	67–232
TaHSP90B-7B	*TraesCS7B01G446900*	15	13	24–66	65–203
TaHSP90B-7D	*TraesCS7D01G517800*	15	7	20–78	65–235
TaHSP90C1-5A	*TraesCS5A01G251000*	19	2	32–89	29–190
TaHSP90C1-5B	*TraesCS5B01G249000*	19	4	22–119	77–422
TaHSP90C1-5D	*TraesCS5D01G258900*	19	2	15–99	27–220
TaHSP90C2-5A	*TraesCS5A01G101900*	20	5	5–12	4–24
TaHSP90C2-5B	*TraesCS5B01G106300*	20	9	6–12	19–104
TaHSP90C2-5D	*TraesCS5D01G113700*	20	7	7–13	11–56

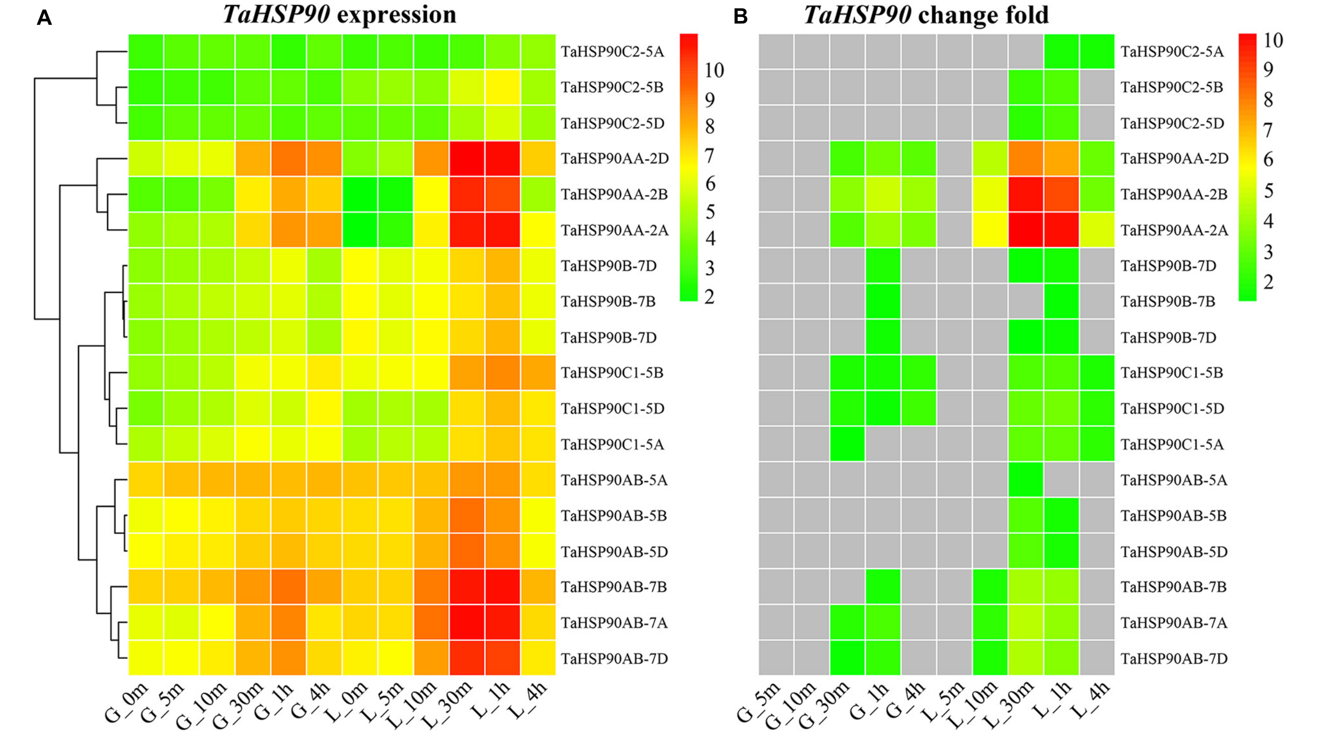


FIGURE 3 | Expression and response patterns of *TaHSP90s* at different heat stress treatment (37°C) time points. The samples were indicated by a character that refers to the organ, followed by the heat stress time point (for example, "G_5m" indicated that grains were treated with heat stress for 5 m). G, grains; L, flag leaves; m, minute; h, hour. **(A)** Expression level of each *TaHSP90* in 12 samples. The heatmap was drawn by the "pheatmap" package in R software based on log2-transformed (FPKM + 1) value. Complete linkage clustering analysis was performed with maximum distance. **(B)** Change fold of each *TaHSP90* under heat stress. Change fold was calculated by comparing the expression abundance in each heat stress condition to that in normal condition (G- or L-0m). Log2-transformed change fold was shown.

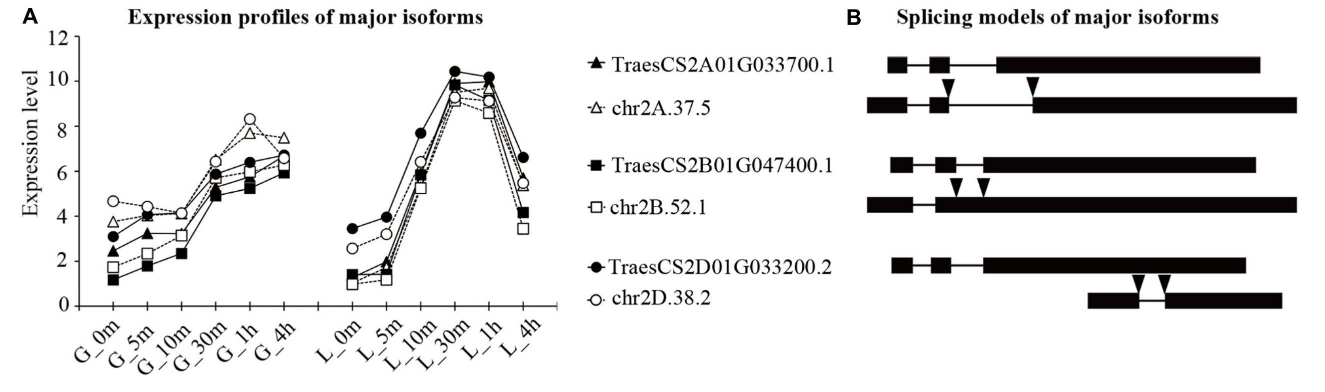


FIGURE 4 | Expression profiles and splicing modes of major isoforms generated by three *TaHSP90AA* homeologs. Isoform names are shown for both **(A)** and **(B)**. **(A)** Expression profiles of major isoforms of *TaHSP90AAs*. Expression level is shown as the log2-transformed (FPKM + 1) value. The samples are indicated by a character that refers to the organ, followed by the heat stress duration (for example, "G_5m" indicated that grains were treated with heat stress for 5 m). G, grains; L, flag leaves; m, minute; h, hour. **(B)** Splicing modes of the major isoforms of *TaHSP90AAs*. Solid boxes represented exons, lines represented introns, and arrows indicated the splicing sites that differed from the longest intact coding sequence.

normal conditions in grains and flag leaves, and they were sharply upregulated (fold change ≥ 2 and FDR-adjusted *P-value* < 0.01) at 10 and 30 min heat stress treatment point in flag leaves, and grains, respectively (**Figure 3B**), consistent with *HSP90AAs* that were highly heat inducible (Swindell et al., 2007; Hu et al., 2009). Besides, all of the 18 *TaHSP90s* were heat responsive in flag leaves; however, it is also worth noting that all *TaHSP90C2s* (*TaHSP90C2-5A*, *TaHSP90C2-5B*, and *TaHSP90C2-5D*) and three *TaHSP90ABs* homeologs (*TaHSP90AB-5A*, *TaHSP90AB-5B*, and *TaHSP90AB-5D*) did not respond to heat stress in grains, suggesting a distinct response network between these two organs. In conclusion, together with the high level of sequence and gene structure conservation, these results demonstrated that no significant divergence of the three *TaHSP90* homeologs occurred at transcriptional level in hexaploid wheat, though the heat response patterns of *TaHSP90s* were highly dynamic between different heat durations, gene subfamilies, and organs.

Large Number of Novel Isoforms Generated by TaHSP90s Under Heat Stress

Recent findings have suggested the importance of AS regulation in abiotic stress response (Reddy et al., 2013; Jiang et al., 2017; Keller et al., 2017; Laloum et al., 2018; Liu et al., 2018). Using the qualitative and quantitative heat response transcriptomes of filling grains and flag leaves produced by combining second- and third-generation sequencing in our previous study (Wang et al., 2019), we comprehensively investigated the roles of AS in the heat response of *TaHSP90s*.

First, a total of 126 isoforms of *TaHSP90s* were identified from our data, including the 36 isoforms that had been annotated in IWGSC RefSeq v1.0 and 90 newly identified isoforms (**Supplementary Tables 3**, **4** and **Supplementary Figures 4**, **5**). The number of isoforms per *TaHSP90* gene ranged from 2 to 13, with an average of 9.0, 8.5, 7.0, 6.3, and 2.7 for the subfamily B, AB, C2, AA, and C1, respectively (**Table 1**). Although the exon–intron structures were highly conserved among the three *TaHSP90* homeologs (**Supplementary Figure 2B**), the isoform numbers generated by the three *TaHSP90* homeologs were such distinct. The most distinct change was observed in the three *TaHSP90AB* homeologs on chromosome group 7; the isoform numbers ranged from 6 to 13. In this case, the highly conserved *TaHSP90s* homeologs would possibly diverge at AS level by generating different isoform numbers under heat stress.

Second, with the quantified information of each isoform, we found that the six *TaHSP90s* (*TaHSP90AB-5A*, *TaHSP90AB-5B*, *TaHSP90AB-5D*, *TaHSP90C2-5A*, *TaHSP90C2-5B*, and *TaHSP90C2-5D*) that were not heat responsive in grains generated some isoforms that responded to heat stress with transcriptional regulation (fold change ≥ 2 and FDR-adjusted *P-value* < 0.01) (**Supplementary Figure 6**). Thus, these transcriptionally heat-responsive isoforms extended our understanding of the transcriptional regulation of *TaHSP90s* and further revealed the complexity of the heat stress response for this gene family.

Next, to characterize the predominant isoforms of each *TaHSP90* gene that may play more important roles, we introduced the IEP, which was calculated as the expression abundance ratio of one isoform to all isoforms generated by the same gene. An isoform with an average IEP of more than 30% across all of the time points in an organ was regarded as a major isoform, an isoform with an IEP less than 5% in all time points was regarded as a rare isoform, and all other isoforms were classified as minor isoforms. This analysis led to the classification of isoforms into major (30), minor (44), and rare (52) isoforms (**Supplementary Table 5**). For 18 *TaHSP90s*, one *TaHSP90* (*TaHSP90AB-5B*) generated three major isoforms, 12 *TaHSP90s* generated two major isoforms, and five *TaHSP90s* generated one major isoform. Interestingly, among the two or three major isoforms, one was already annotated in IWGSC RefSeq v1.0 and contained the longest complete coding region, which potentially encoded a functional peptide containing both the HATPase domain and HSP90 domain. However, another

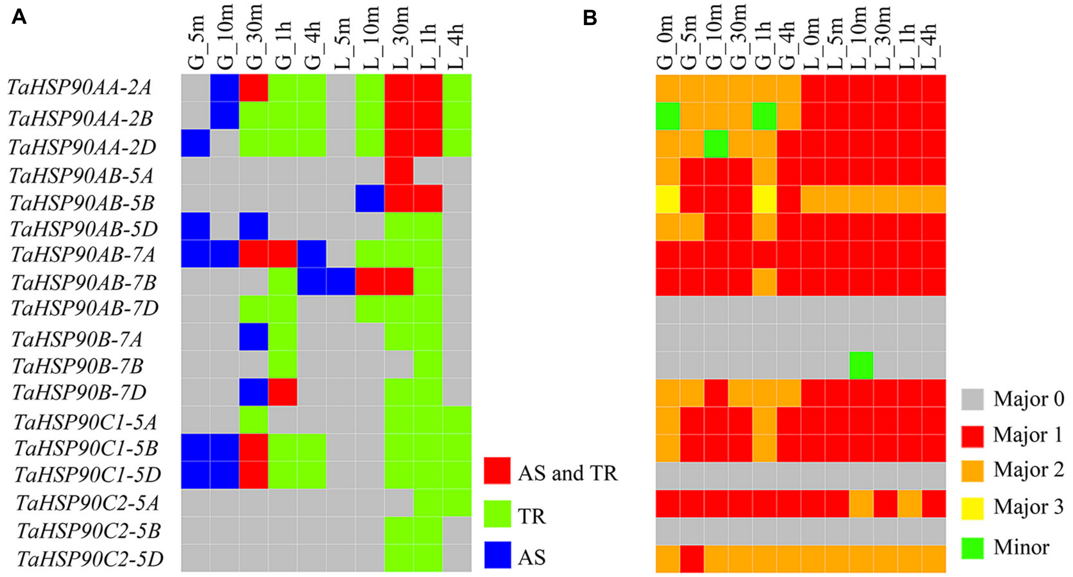


FIGURE 5 | Differential AS response of *TaHSP90s* under heat stress. The samples are indicated by a character that refers to the organ, followed by the heat stress duration (for example, "G_5m" indicated that grains were treated with heat stress for 5 m). G, grains; L, flag leaves; m, minute; h, hour. **(A)** Responses of *TaHSP90s* at different levels under different conditions. AS, alternative splicing response. TR, transcriptional response. **(B)** The most abundant isoform of each *TaHSP90* in each sample. For the five *TaHSP90s* (*TaHSP90AB-7D*, *TaHSP90B-7A*, *TaHSP90B-7B*, *TaHSP90C1-5D*, *TaHSP90C2-5B*), they only generated one major isoform (Major 0). For the other *TaHSP90s*, which generated more than one major isoform; Major 1 represented the major isoform with the complete reading frame, and Major 2 and Major 3 (for *TaHSP90AB-5B* only) represented the other major isoforms. Minor represented the minor isoform.

major isoform was newly discovered from our hybrid sequencing data, and this possessed only the HSP90 domain that potentially encoded truncated peptides. Furthermore, expression analysis showed that the major isoforms generated by the same gene had comparable expression levels and response patterns (**Figure 4** and **Supplementary Figures 7**, **8**), making it intriguing as to what roles these newly discovered isoforms played in the heat stress response.

Varied Number and Splicing Modes of Major Isoforms Generated by TaHSP90 Homeologs

The above transcription analysis showed that the expression abundance and response patterns of three *TaHSP90* homeologs were conserved. However, regarding the number of major isoforms generated by three *TaHSP90* homeologs, only one or two homeologs generated two major isoforms in subfamily AB, B, C1, and C2, with the exception of subfamily AA in which each *TaHSP90* homeolog generated two major isoforms (**Figure 4** and **Supplementary Figures 7**, **8**), demonstrating a varied major isoform number among three *TaHSP90* homeologs.

Furthermore, the newly identified major isoforms (NIMIs) from our hybrid sequencing data also exhibited different exon and intron compositions among the three *TaHSP90* homeologs, although they had the same gene structure, implying that the three *TaHSP90* homeologs may exploit different peptides to respond to heat stress. For example, among the three *TaHSP90AAs* homeologs, the NIMI of *TaHSP90AA-2A* underwent AS at both the 5′ (Alt5'SS) and the 3′ ends of the first intron (Alt3'SS). The NIMI of *TaHSP90AA-2B* underwent intron retention. The NIMI of *TaHSP90AA-2D* underwent exon skipping (**Figure 4**). In conclusion, the number and isoform structure generated by the three *TaHSP90* homeologs were significantly different, despite the three *TaHSP90* homeologs being conserved at the sequence level and at the transcriptional response level. The differential number and isoform structure possibly suggested a new direction for evolutionary divergence of *TaHSP90* homeologs in hexaploid wheat.

Differential AS Responses of Three TaHSP90 Homeologs

Comparing the control and heat stress treatment samples, we identified 12 *TaHSP90s*, which responded to heat stress by generating new isoforms or changing the expression level of highly expressed isoforms with the criterion defined in our previous study (**Figure 5A**) (Wang et al., 2019). Interestingly, some *TaHSP90s* that did not respond to heat stress with transcriptional regulation did respond to heat stress with AS regulation, particularly in grains under short time (5 and 10 min) heat stress, extending our understanding of the heat response of *TaHSP90s*. Significantly, for the three *TaHSP90* homeologs in subfamily B, C1, and AB on chromosome group 7, only one or two homeologs responded to heat stress with AS regulation, and in subfamily AA, though all three *TaHSP90AAs* underwent AS responses, but the AS responses occurred at different heat stress time points in grains. These results demonstrated the differential responses of three *TaHSP90s* homeologs at the AS regulation level.

Furthermore, using the qualitative and quantitative isoforms, we investigated the isoform with the highest abundance for each *TaHSP90* at each heat stress time point sample (**Figure 5B**). The highest abundance isoforms generated by three *TaHSP90* homeologs at specific time point were also distinct, providing another evidence for the differential AS responses and suggesting possible diverged evolution. In conclusion, inconsistent with the conserved sequences and transcriptional regulation, AS response diverged among the three *TaHSP90* homeologs, extending our understanding about the functional conservation and divergence of this gene family in hexaploid wheat.

DISCUSSION

HSP90s play vital roles in plant growth and stress response (Wang et al., 2004; Wang et al., 2016; Gil et al., 2017). In this study, we first performed a genome-wide analysis of *HSP90s* in hexaploid wheat and its progenitors. The copy numbers of *HSP90s* among these species were not consistent with polyploidy level. Next, we comprehensively analyzed the heat response patterns of *TaHSP90s* and found that AS diversified the heat response of *TaHSP90s*, suggesting different options for functional studies and breeding strategies. Meanwhile, our results provided a new perspective for understanding about evolutionary conservation and divergence for the homeologous genes in polyploidy species.

It has been reported that *HSP90AAs* were dramatically upregulated during heat stress, and *HSP90ABs* were constitutively expressed in Arabidopsis (Swindell et al., 2007) and rice (Hu et al., 2009). However, with more intensive time course transcriptomes and different organs, we showed that all of the *TaHSP90s* were heat responsive under at least one time point in flag leaves, suggesting that this was a specific feature in hexaploid wheat or hinting that more investigations should be performed in Arabidopsis and rice to draw a conclusion. However, it was noteworthy that our data mainly focused on the short time heat response, and the *TaHSP90s* that do not respond to heat stress in grains may also respond to heat stress in other conditions.

The differentiation and subfunctionalization of homeologous genes are intensified by stress and are thought to contribute to the acclimation of polyploidy plants to stress (Dong and Adams, 2011; Liu et al., 2015; Powell et al., 2017). For example, about 68% of homeologous genes display expression partitioning according to the extent of stress responsiveness in hexaploid wheat (Liu et al., 2015). The homeolog-specific expression patterns of homeologous genes were also widely reported in wheat genes, resulting in different morphological phenotypes like lateral root number (Wang et al., 2018a), root hair length (Han et al., 2016), and other domestication traits (Zhang et al., 2011). For HSP90s, members of three *TaHSP90C2* homeologs and three *TaHSP90AB* homeologs were also predicted to experience expression partitioning under drought stress and combined drought-heat stress, but not under only heat stress (Liu et al., 2015). Similarly, in our analysis, the heat expression trends and fold changes of *TaHSP90* members were not significantly distinguishable between each other in three *TaHSP90* homeologs. Thus, it seemed like the three *TaHSP90* homeologs were conserved and had not undergone subfunctionalization or neofunctionalization in heat response at transcriptional level, which was consistent with the highly conserved sequences and motifs.

However, in our subsequent analysis, we found that the numbers of major isoforms were distinct among three *TaHSP90* homeologs, and further investigations revealed that the AS modes of the major isoforms generated by the three *TaHSP90* homeologs were also not conserved, suggesting that the differentiation of *TaHSP90* homeologs may occur at the AS level. Different splicing patterns have also been characterized among homeologous genes in allopolyploid cotton (Wang et al., 2018b). More than 51% of the homeologous genes generated isoforms containing different structure in allotetraploid cottons (Wang et al., 2018b). Theoretically, the differentiation in AS that resulted in distinct transcripts may lead to the diversified functions of homeologous genes (Long et al., 2013). Thus, divergent AS patterns and differential AS responses may contribute to the functional divergence and differential evolution of *TaHSP90* homeologs, changing our understanding of the conservation of *HSP90s* in terms of expression profile and function.

In this study, *TaHSP90s* were found to generate many novel isoforms in the grain-filling stage under heat stress. Contrary to that, the expression of the abnormal isoforms was generally lower than those of the full-length isoforms of *Lipoxygenase* members in the tea plant in response to low temperature (Zhu et al., 2018), the expression levels of the truncated major isoforms and their full-length counterparts were found to be comparable in the current study. It was worthy to note that the peptides encoded by the truncated major isoforms only contain the HSP90 domain (**Supplementary Table 3**), making the roles of these truncated major isoforms intriguing. It is well known that the HATPase domain is responsible to bind ATP; the HSP90 domain is responsible for homodimerization and binding to clients (Schopf et al., 2017). We proposed that the truncated peptides modify their original functions by modulating the domain composition. For example, the novel major isoform of *TaHSP90AA-2A* possibly encoded a peptide containing the intact HSP90 domain but lost the HATPase domain. The truncated peptide may still form a homodimer but fail to bind clients without the ability to bind ATP; it seemed to decrease the protecting capacity of HSP90, and this was quite different to the AS regulation of *HSFA2* under heat stress (Sugio et al., 2009; Liu et al., 2013; Cheng et al., 2015). In contrast, a total of 70 isoforms, including 36 novel isoforms, were annotated to comprise both the HATPase and HSP90 domain. By this way, these isoforms would probably encode different proteins and significantly increase HSP90 protein diversity to protect different substrates. The contrary hypothesis of the roles of the different isoforms should be elucidated in further studies. Furthermore, as HSP90s were found to translocate into the nucleus under heat stress, alternative TaHSP90 isoforms finally were transported into the nucleus would result in a different expression of different heat-responsive genes. Another question was how many isoforms could be finally translated into proteins, as isoforms arose from AS always subsequently translated into normal or truncated proteins, or degraded by the non-sense-mediated decay pathway (Kalyna et al., 2012; Syed et al., 2012; Chaudhary et al., 2019)?

In grains, the *TaHSP90AAs*, the major heat-responsive *HSP90s*, tended to favor the novel truncated major isoform and minor isoform in most samples; the other eight *TaHSP90s* also favored different isoforms in different samples. These changes raised questions whether they correlated to the delay heat response in grains or thermotolerance of grains. These results also remind us that when investigating the functions of homeologous genes, expressions and splicing types of isoforms would be important and worthy of study.

About 40% of the differentially spliced genes were also found to be regulated at the transcriptional level, inferring the vital role of the cooperation of AS and transcriptional regulation in heat response in hexaploid wheat (Liu et al., 2018). In the present study, all of the 18 *TaHSP90s* were transcriptionally regulated, and 12 of these were also AS regulated. The higher cooperation ratio of AS and transcriptional regulation may contribute to fine modulation of *TaHSP90s*, to match its key roles in heat stress response.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/sra/SRP128236.

AUTHOR CONTRIBUTIONS

SX and XW designed the study. YL, PZ, AZ, LM, SX, and XW analyzed the data. YL, SX, and XW wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Key Research and Development Program of China (2016YFD0101602), the National Natural Science Foundation of China (31501380), and the Open Project Program of the State Key Laboratory of Crop Stress Biology for Arid Areas (CSBAA2019005).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Research Square (Lu et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.577897/full#supplementary-material

REFERENCES

Agarwal, G., Garg, V., Kudapa, H., Doddamani, D., Pazhamala, L. T., Khan, A. W., et al. (2016). Genome-wide dissection of AP2/ERF and HSP90 gene families in five legumes and expression profiles in chickpea and pigeonpea. *Plant Biotechnol. J.* 14, 1563–1577. doi: 10.1111/pbi.12520

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nat. Clim. Change* 5:143.

Borrill, P., Adamski, N., and Uauy, C. (2015). Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol.* 208, 1008–1022. doi: 10.1111/nph.13533

Carretero-Paulet, L., Albert, V. A., and Fares, M. A. (2013). Molecular evolutionary mechanisms driving functional diversification of the HSP90A family of heat shock proteins in eukaryotes. *Mol. Biol. Evol.* 30, 2035–2043. doi: 10.1093/molbev/mst113

Chaudhary, S., Jabre, I., Reddy, A. S. N., Staiger, D., and Syed, N. H. (2019). Perspective on alternative splicing and proteome complexity in plants. *Trends Plant Sci.* 24, 496–506. doi: 10.1016/j.tplants.2019.02.006

Chen, B., Zhong, D., and Monteiro, A. (2006). Comparative genomics and evolution of the HSP90 family of genes across all kingdoms of organisms. *BMC Genom.* 7:156. doi: 10.1186/1471-2164-7-156

Cheng, Q., Zhou, Y., Liu, Z., Zhang, L., Song, G., Guo, Z., et al. (2015). An alternatively spliced heat shock transcription factor, OsHSFA2dI, functions in the heat stress-induced unfolded protein response in rice. *Plant Biol.* 17, 419–429. doi: 10.1111/plb.12267

Dong, S., and Adams, K. L. (2011). Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. *New Phytol.* 190, 1045–1057. doi: 10.1111/j.1469-8137.2011.03650.x

Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316, 1862–1866. doi: 10.1126/science.1143986

Geiler-Samerotte, K. A., Zhu, Y. O., Goulet, B. E., Hall, D. W., and Siegal, M. L. (2016). Selection transforms the landscape of genetic variation interacting with Hsp90. *PLoS Biol.* 14:e2000465. doi: 10.1371/journal.pbio.2000465

Gil, K.-E., Kim, W.-Y., Lee, H.-J., Faisal, M., Saquib, Q., Alatar, A. A., et al. (2017). ZEITLUPE contributes to a thermoresponsive protein quality control system in *Arabidopsis*. *Plant Cell* 29, 2882–2894. doi: 10.1105/tpc.17.00612

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hahn, A., Bublak, D., Schleiff, E., and Scharf, K.-D. (2011). Crosstalk between Hsp90 and Hsp70 chaperones and heat stress transcription factors in tomato. *Plant Cell* 23, 741–755. doi: 10.1105/tpc.110.076018

Han, Y., Xin, M., Huang, K., Xu, Y., Liu, Z., Hu, Z., et al. (2016). Altered expression of TaRSL4 gene by genome interplay shapes root hair length in allopolyploid wheat. *New Phytol.* 209, 721–732. doi: 10.1111/nph.13615

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hu, W., Hu, G., and Han, B. (2009). Genome-wide survey and expression profiling of heat shock proteins and heat shock factors revealed overlapped and stress specific response under abiotic stresses in rice. *Plant Sci.* 176, 583–590. doi: 10.1016/j.plantsci.2009.01.016

International Wheat Genome Sequencing Consortium [IWGSC] (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Jiang, J., Liu, X., Liu, C., Liu, G., Li, S., and Wang, L. (2017). Integrating omics and alternative splicing reveals insights into grape response to high temperature. *Plant Physiol.* 173, 1502–1518. doi: 10.1104/pp.16.01305

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587. doi: 10.1038/nmeth.4285

Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., et al. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* 40, 2454–2469. doi: 10.1093/nar/gkr932

Keller, M., Hu, Y., Mesihovic, A., Fragkostefanakis, S., Schleiff, E., and Simm, S. (2017). Alternative splicing in tomato pollen in response to heat stress. *DNA Res.* 24, 205–217.

Kolde, R., and Kolde, M. R. (2015). *Package 'Pheatmap'*. http://cran.r-project.org/web/packages/pheatmap/index.html (accessed March 12, 2020).

Krishna, P., and Gloor, G. (2001). The Hsp90 family of proteins in *Arabidopsis thaliana*. *Cell Stress Chaperones* 6:238. doi: 10.1379/1466-1268(2001)006<0238:thfopi>2.0.co;2

Laloum, T., Martín, G., and Duque, P. (2018). Alternative splicing control of abiotic stress responses. *Trends Plant Sci.* 23, 140–150. doi: 10.1016/j.tplants.2017.09.019

Leach, L. J., Belfield, E. J., Jiang, C., Brown, C., Mithani, A., and Harberd, N. P. (2014). Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. *BMC Genom.* 15:276. doi: 10.1186/1471-2164-15-276

Lesk, C., Rowhani, P., and Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature* 529, 84–87. doi: 10.1038/nature16467

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493–D496.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259.

Liu, B., Liu, L., Tian, L., Cao, W., Zhu, Y., and Asseng, S. (2014). Post-heading heat stress and yield impact in winter wheat of China. *Global Change Biol.* 20, 372–381. doi: 10.1111/gcb.12442

Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., et al. (2013). An autoregulatory loop controlling *Arabidopsis* HsfA2 expression: role of heat shock-induced alternative splicing. *Plant Physiol.* 162, 512–521. doi: 10.1104/pp.112.205864

Liu, Z., Qin, J., Tian, X., Xu, S., Wang, Y., Li, H., et al. (2018). Global profiling of alternative splicing landscape responsive to drought, heat and their combination in wheat (Triticum aestivum L.). *Plant Biotechnol. J.* 16, 714–726. doi: 10.1111/pbi.12822

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., et al. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum L.). *BMC Plant Biol.* 15:152. doi: 10.1186/s12870-015-0511-8

Long, M., VanKuren, N. W., Chen, S., and Vibranovski, M. D. (2013). New gene evolution: little did we know. *Ann. Rev. Genet.* 47, 307–333. doi: 10.1146/annurev-genet-111212-133301

Lu, Y., Zhao, P., Zhang, A., Ma, L., Xu, S., and Wang, X. (2019*). Alternative splicing diversifies the heat response and evolutionary strategy of conserved Heat Shock Protein 90 in bread wheat (Triticum aestivum L.). *Res. Square* doi: 10.21203/rs.2.17636/v1

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 345:1250092.

Meiri, D., and Breiman, A. (2009). *Arabidopsis* ROF1 (FKBP62) modulates thermotolerance by interacting with HSP90.1 and affecting the accumulation of HsfA2-regulated sHSPs. *Plant J.* 59, 387–399. doi: 10.1111/j.1365-313x.2009.03878.x

Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi: 10.1093/molbev/mst024

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Ohama, N., Sato, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2017). Transcriptional regulatory network of plant heat stress response. *Trends Plant Sci.* 22, 53–65. doi: 10.1016/j.tplants.2016.08.015

Powell, J. J., Fitzgerald, T. L., Stiller, J., Berkman, P. J., Gardiner, D. M., Manners, J. M., et al. (2017). The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnol. J.* 15, 533–543. doi: 10.1111/pbi.12651

Prasad, P. V., and Djanaguiraman, M. (2014). Response of floret fertility and individual grain weight of wheat to high temperature stress: sensitive stages and thresholds for temperature and duration. *Funct. Plant Biol.* 41, 1261–1269. doi: 10.1071/fp14061

Queitsch, C., Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature* 417, 618–624. doi: 10.1038/nature749

Ramírez-González, R., Borrill, P., Lang, D., Harrington, S., Brinton, J., Venturini, L., et al. (2018). The transcriptional landscape of polyploid wheat. *Science* 361:eaar6089.

Reddy, A. S., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* 25, 3657–3683. doi: 10.1105/tpc.113.117523

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rutherford, S. L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342. doi: 10.1038/24550

Schopf, F. H., Biebl, M. M., and Buchner, J. (2017). The HSP90 chaperone machinery. *Nat. Rev. Mol. Cell Biol.* 18:345.

Song, Z., Pan, F., Yang, C., Jia, H., Jiang, H., He, F., et al. (2019). Genome-wide identification and expression analysis of HSP90 gene family in Nicotiana tabacum. *BMC Genet.* 20:35. doi: 10.1186/s12863-019-0738-8

Sugio, A., Dreos, R., Aparicio, F., and Maule, A. J. (2009). The cytosolic protein response as a subcomponent of the wider heat shock response in *Arabidopsis*. *Plant Cell* 21, 642–654. doi: 10.1105/tpc.108.062596

Swindell, W. R., Huebner, M., and Weber, A. P. (2007). Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genom.* 8:125. doi: 10.1186/1471-2164-8-125

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J. W. (2012). Alternative splicing in plants–coming of age. *Trends Plant Sci.* 17, 616–623. doi: 10.1016/j.tplants.2012.06.001

Wang, G. F., Wei, X., Fan, R., Zhou, H., Wang, X., Yu, C., et al. (2011). Molecular analysis of common wheat genes encoding three types of cytosolic heat shock protein 90 (Hsp90): functional involvement of cytosolic Hsp90s in the control of wheat seedling growth and disease resistance. *New Phytol.* 191, 418–431. doi: 10.1111/j.1469-8137.2011.03715.x

Wang, H., Hu, Z., Huang, K., Han, Y., Zhao, A., Han, H., et al. (2018a). Three genomes differentially contribute to the seedling lateral root number in allohexaploid wheat: evidence from phenotype evolution and gene expression. *Plant J.* 95, 976–987. doi: 10.1111/tpj.14005

Wang, M., Wang, P., Liang, F., Ye, Z., Li, J., Shen, C., et al. (2018b). A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol.* 217, 163–178. doi: 10.1111/nph.14762

Wang, R., Zhang, Y., Kieffer, M., Yu, H., Kepinski, S., and Estelle, M. (2016). HSP90 regulates temperature-dependent seedling growth in *Arabidopsis* by stabilizing the auxin co-receptor F-box protein TIR1. *Nat. Commun.* 7:10269.

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* 9, 244–252. doi: 10.1016/j.tplants.2004.03.006

Wang, X., Chen, S., Shi, X., Liu, D., Zhao, P., Lu, Y., et al. (2019). Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat. *Plant J.* 98, 1015–1032. doi: 10.1111/tpj.14299

Xu, Z.-S., Li, Z.-Y., Chen, Y., Chen, M., Li, L.-C., and Ma, Y.-Z. (2012). Heat shock protein 90 in plants: molecular mechanisms and roles in stress responses. *Int. J. Mol. Sci.* 13, 15706–15723. doi: 10.3390/ijms131215706

Yabe, N., Takahashi, T., and Komeda, Y. (1994). Analysis of tissue-specific expression of *Arabidopsis thaliana* HSP90-family gene HSP81. *Plant Cell Physiol.* 35, 1207–1219. doi: 10.1093/oxfordjournals.pcp.a078715

Zhang, J., Li, J., Liu, B., Zhang, L., Chen, J., and Lu, M. (2013). Genome-wide analysis of the Populus Hsp90 gene family reveals differential expression patterns, localization, and heat stress responses. *BMC Genom.* 14:532. doi: 10.1186/1471-2164-14-532

Zhang, M., Shen, Z., Meng, G., Lu, Y., and Wang, Y. (2017). Genome-wide analysis of the *Brachypodium distachyon* (L.) P. Beauv. Hsp90 gene family reveals molecular evolution and expression profiling under drought and salt stresses. *PLoS One* 12:e0189187. doi: 10.1371/journal.pone.0189187

Zhang, Y., Liu, Z., Khan, A. A., Lin, Q., Han, Y., Mu, P., et al. (2016). Expression partitioning of homeologs and tandem duplications contribute to salt tolerance in wheat (Triticum aestivum L.). *Sci. Rep.* 6:21476.

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., et al. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proc. Natl. Acad. Sci. U S A.* 108, 18737–18742. doi: 10.1073/pnas.1110552108

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci. U S A.* 114, 9326–9331. doi: 10.1073/pnas.1701762114

Zhu, J., Wang, X., Guo, L., Xu, Q., Zhao, S., Li, F., et al. (2018). Characterization and alternative splicing profiles of the Lipoxygenase gene family in tea plant (Camellia sinensis). *Plant Physiol.* 59, 1765–1781. doi: 10.1093/pcp/pcy091

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/fgene.2020.603056

Excision Dominates Pseudogenization During Fractionation After Whole Genome Duplication and in Gene Loss After Speciation in Plants

Zhe Yu[1], *Chunfang Zheng*[1], *Victor A. Albert*[2] *and David Sankoff*[1]*

[1] *Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada,* [2] *Department of Biological Sciences, University at Buffalo, Buffalo, NY, United States*

We take advantage of synteny blocks, the analytical construct enabled at the evolutionary moment of speciation or polyploidization, to follow the independent loss of duplicate genes in two sister species or the loss through fractionation of syntenic paralogs in a doubled genome. By examining how much sequence remains after a contiguous series of genes is deleted, we find that this residue remains at a constant low level independent of how many genes are lost—there are few if any relics of the missing sequence. Pseudogenes are rare or extremely transient in this context. The potential exceptions lie exclusively with a few examples of speciation, where the synteny blocks in some larger genomes tolerate degenerate sequence during genomic divergence of two species, but not after whole genome doubling in the same species where fractionation pressure eliminates virtually all non-coding sequence.

Keywords: gene loss, fractionation, polyploidization, whole genome duplication, plant evolution, synteny, pseudogene, genomics

OPEN ACCESS

Edited by:
Chuang Ma,
Northwest A and F University, China

Reviewed by:
Vladimir A. Trifonov,
Institute of Molecular and Cellular Biology (RAS), Russia
Weilong Hao,
Wayne State University, United States

****Correspondence:***
David Sankoff
sankoff@uottawa.ca

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *04 September 2020*
Accepted: *27 November 2020*
Published: *18 December 2020*

Citation:
Yu Z, Zheng C, Albert VA and Sankoff D (2020) Excision Dominates Pseudogenization During Fractionation After Whole Genome Duplication and in Gene Loss After Speciation in Plants. Front. Genet. 11:603056. doi: 10.3389/fgene.2020.603056

1. INTRODUCTION

The evolutionary process of gene loss, through DNA excision—or sequence elimination (Eckardt, 2001), pseudogenization (Jacq et al., 1977), or other mechanism, is the obverse of gene acquisition by a genome through processes such as tandem or remote duplication of individual genes, whole genome doubling (WGD), neo- and sub-functionalization and horizontal transfer. Loss serves a number of functional and structural roles, such as in the reconfiguring of regulatory or metabolic networks or in compensating for the energetic, material, and structural costs of gene complement expansion.

An longstanding biological controversy in evolutionary genomics (Byrnes et al., 2006; van Hoek and Hogeweg, 2007) involves the question of whether duplicated genes are deleted through random excision "elimination of excess DNA" namely the deletion of chromosomal segments containing one or more genes, which we have termed the "structural" mechanism, or through targeted (possibly) gene-by gene events such as regulatory epigenetic silencing and pseudogenization, which we call "functional" mechanisms. Because it is often difficult to ascertain whether a single-copy gene is the result of the deletion of a duplicate copy, and because the outcomes of the two kinds of process may appear similar, it is often difficult to discern which one is operating.

The alignment of the gene orders of homologous genes in two related genomes, or subgenomes of an (ancient) polyploid, such as that provided by the SynMap program on the CoGe platform (Lyons and Freeling, 2008; Lyons et al., 2008), is a uniquely reliable first step in the assessment of gene conservation or loss after speciation or polyploidization. The homology of pairs of genes in the chromosomal fragments "synteny blocks" making up such an alignment, is doubly confirmed, first by the common level of sequence similarity of all the gene pairs in the block, and second by the common chromosomal context, namely the common order of the homologous genes in the two fragments, represented as follows:

● ○ ● ● ○ ○ ● ● ○ ●
● ● ● ● ● ● ○ ● ● ●

Synteny block on homeologous regions of two chromosomes.
Dark circles indicate retained genes, white circles deleted genes.
There are five retained duplicate gene pairs, four singletons on the lower chromosome and one singleton on the upper chromosome.

In synteny blocks, it is relatively easy to see where duplicate genes have been deleted, and how many genes in a row have been lost. In this paper, we use this property of synteny blocks in devising a simple method to distinguish clearly between genomes where excision is the main mechanism for gene loss, and those where pseudogenization may also play a role.

Although the basics of polyploidy in plants have been understood for over a century (Winge, 1917), and though this process is well-attested across the entire evolutionary spectrum, from bacteria (Hansen, 1978; Tobiason and Seifert, 2006) to pre-mammalian vertebrates (Ohno, 1970), the statistical study of conservation and reduction at the genome level originates with the discovery and analysis by Wolfe and Shields of an ancient WGD in the *Saccharomyces cerevisiae* genome sequence (Wolfe and Shields, 1977). But starting with the first few plant genomes to be sequenced—*Arabidopsis, Oryza, Populus*—the realization has grown that all flowering plants species are "paleopolyploids," re-diploidized descendants of one or more ancient polyploidization events. It is in the context of the Angiosperm/Magnoliophyte phylum or division that we have attempted to resolve the structure-function controversy (Byrnes et al., 2006; van Hoek and Hogeweg, 2007) using several modeling and statistical approaches (Zheng et al., 2009; Sankoff et al., 2010, 2015; Yu and Sankoff, 2016; Yu et al., 2020). In the present paper, however, our focus is less on how fractionated gene pairs are organized within synteny blocks, than on what happens to these genes—do they degenerate in place, or are they simply removed from the DNA sequence of the genome?

Our claim is that the overwhelming loss process is the latter: the complete excision of the gene from the genome, the elimination of the sequence of the entire gene. As such, we do not adopt any restrictive definition of a pseudogene or quantification of the various types of pseudogenes in plants, which was done in the recent definitive study of Xie et al. (2019); here we simply examine whether any DNA, and how much, remains, when a one member of a pair of homeologous genes, as identified by SynMap, is absent from a syntenic block. We will show that in the large majority of cases, there is a drastic loss of DNA, leaving only a small stretch of intergenic sequence, so that no kind of pseudogene, whatever its definition, except for very small fragments of cDNA, can be present. In other words, fractionation, and most gene loss in ancient genomes, does not tend to result in long-lasting full length or part length degenerate genes, but a relatively complete loss of the DNA. This does not mean that pseudogenes are absent or even rare in these and other genomes. Many of these may persist over many millions of years. Nevertheless, Xie et al. (2019) found that poplar has almost 25,000 pseudogenes, but <1,500 of these stem from the Salix whole genome doubling, and most of these are presumably small fragments of coding sequence.

2. METHODS

2.1. Sampling of Plant Species

In each of four core eudicot plant families (or orders), we selected a pair of genomes for which annotated genome sequences are available:

1. *Populus trichocarpa* (poplar) CoGe ID 25127, and *Salix purpurea* (willow) CoGe ID 52439 in the rosid family Salicaceae,
2. *Salvia splendens* (scarlet sage) CoGe ID 55705, and *Tectona grandis* (teak) CoGe ID 55706 in the asterid family Lamiaceae,
3. *Linum usitatissimum* (flax) CoGe ID 16772 and *Hevea brasiliensis* (rubber tree) CoGe ID 16772 in the order Malpighiales, also rosids, and
4. *Malus domestica* (apple) CoGe ID 54783 and *Pyrus* × *bretschneideri* (pear) CoGe ID 37224 belonging to the same subtribe Malinae of another rosid family Rosaceae.

All these genomes have undergone at least one whole genome duplication since the ancient whole genome triplication "gamma" at the origin of the core eudicots.

2.2. Construction of Synteny Blocks

For each of the eight genomes individually we first carried out a self-comparison of the unmasked sequences using the SynMap program on the CoGe platform (Lyons and Freeling, 2008; Lyons et al., 2008) to construct paralogous syntenic blocks. Based on the distribution of gene pair similarities, also output by SynMap, we retained only those blocks for which the average similarity confirmed that the duplication occurred at the time of the most recent polyploidization event experienced by the genome.

For each of the four pairs of genomes, we then used SynMap to compare the two and construct orthologous synteny blocks. We again referred to the distribution of gene pair similarities in selecting only those blocks likely to have been created at the time of the speciation event at the origin of the diverging lineages leading to the two species being studied. We thus aimed to exclude synteny blocks created by polyploidization in the common ancestor of the two, including the gamma triplication, as well as blocks created in either of the two genomes by post-speciation polyploid events.

The stringent criteria, such as a minimum number of contiguous pairs (default = 5), incorporated in SynMap tends

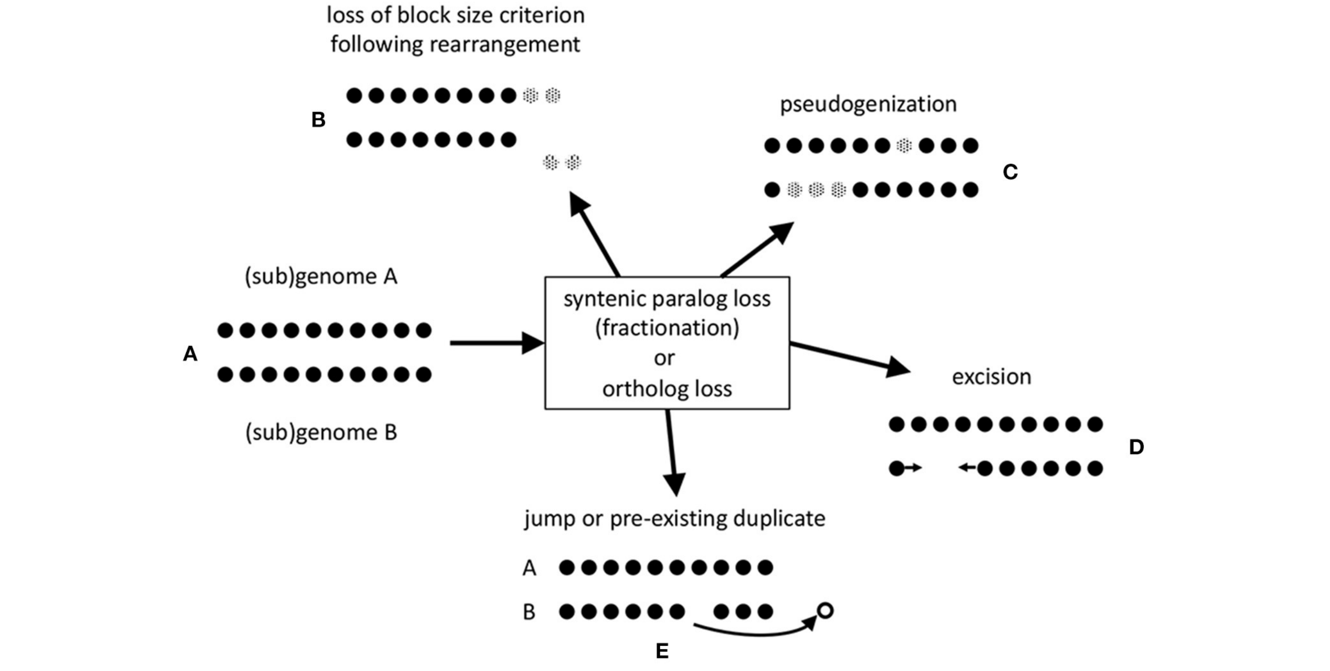


FIGURE 1 | Mechanisms of loss of gene pair homology after speciation or polyploidization. **(A)** Synteny block made up of co-linear homologous pairs. **(B)** erosion of synteny block by translocation to a remote chromosomal location of a portion of sub-threshold length. **(C)** Pseudogenization. Genes rendered inoperable represented by gray dots. **(D)** Excision of DNA fragment including one or more genes. Arrows represent a new adjacency after the loss of the excised genes. **(E)** Jump of one member of pair to a different genomic location, or loss of only one of two or more homologs of the same gene.

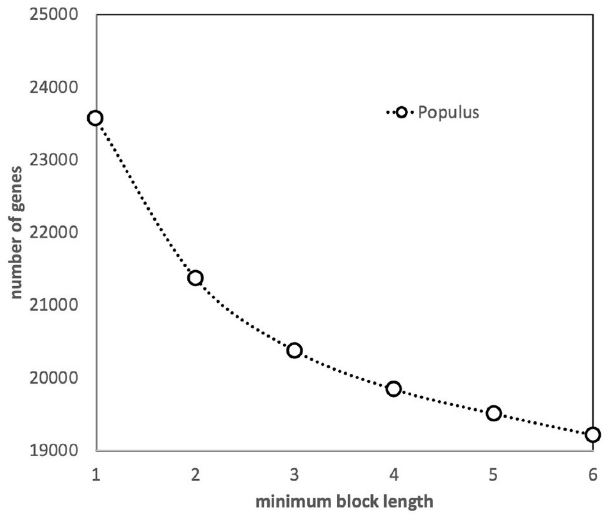


FIGURE 2 | Effect of minimum block size (number of genes) on the number of genes incorporated into synteny blocks.

to excludes some of the homologous gene pairs created by these genomic events (represented in **Figure 1A**), especially after some time has elapsed. Inversions, translocations and other chromosomal rearrangement events in a genome or in either of two related genomes, break synteny blocks into smaller pieces that may not satisfy the criteria, as illustrated in **Figures 1B,E**.

We have assessed the effect of the default SYNMAP requirement—at least five closely spaced gene pairs for a synteny block to be identified—by increasing and decreasing this threshold (see **Figure 2**). A slight decrease in the number of genes in blocks when the threshold is increased to 6 is simply due to the elimination of a few blocks of length 5. But as we decrease the threshold to 3, the algorithm starts to capture blocks made up of independently created but coincidentally neighboring pairs, as well as pairs where one member is already in a larger block, since a gene can be in more than one block. It becomes increasingly difficult to disentangle the behavior of duplicate gene pairs created by polyploidization from other processes of duplication and loss. Thus, we retained the default value, 5.

Since we will be focusing on pseudogenization and excision in our analysis, **Figures 1C,D**, we developed a method that does not favor the identification of one in favor of the other.

2.3. Identification of Deletion Intervals and Their Lengths

We scanned the output of the retained synteny blocks for homeologous segments on two chromosomes (or two disjoint regions of one chromosome) bounded by one or (usually) more duplicate gene pairs at both ends, where all the genes in one segment—the fractionated side—are absent, i.e., not detected by SYNMAP (No gene can be absent from the other segment—otherwise the ancient gene pair, if it ever existed, would not be visible.) We call the number of contiguous single-copy genes in the unfractionated side of the segment the *length* of the interval.

This is the same as the number of genes that are missing from the fractionated side.

For both sides of the segment, we also determine the amount of DNA between the pairs that bound the segment. For the unfractionated side, with all the single-copy genes, this is just the size (in base pairs) of the genes plus the intergenic regions, including the initial region, after one bounding pair, and the final region, before the other bounding pair, in the segment. In the fractionated side, this includes whatever DNA remains between the two bounding pairs, which does not include any genes, according to SynMap.

Two possibilities are represented by **Figures 1C,D**. In the former case, pseudogenization, a gene is rendered inoperable, such as by a point mutation that creates a stop codon inside an erstwhile coding region. In the latter, a chromosomal fragment containing one or more genes is simply physically excised. To assess which of these two processes accounts for the data, we note that pseudogenization through acquiring a gene-internal stop codon, or a frameshift, leaving the gene intact, at least initially, does not shorten the length of the chromosomal region it is in. The average length of a pseudogene is roughly half of that of a functional gene (Xie et al., 2019), but this average includes the very numerous short fragments. In contrast, excision of genes, including some or all of the flanking intergenic DNA, will definitely shorten the region, leaving at most a short stretch of non-coding sequence.

2.4. The Visualization of Gene Density and Pseudogene Density

By plotting the average number of base-pairs in the unfractionated, or totally conserved, intervals of a given length against the length of the interval, we estimate the average size of a gene (plus the following intergenic region). In most cases we expect this plot to be approximately linear, with slope giving the average base-pairs per gene. This is just the inverse

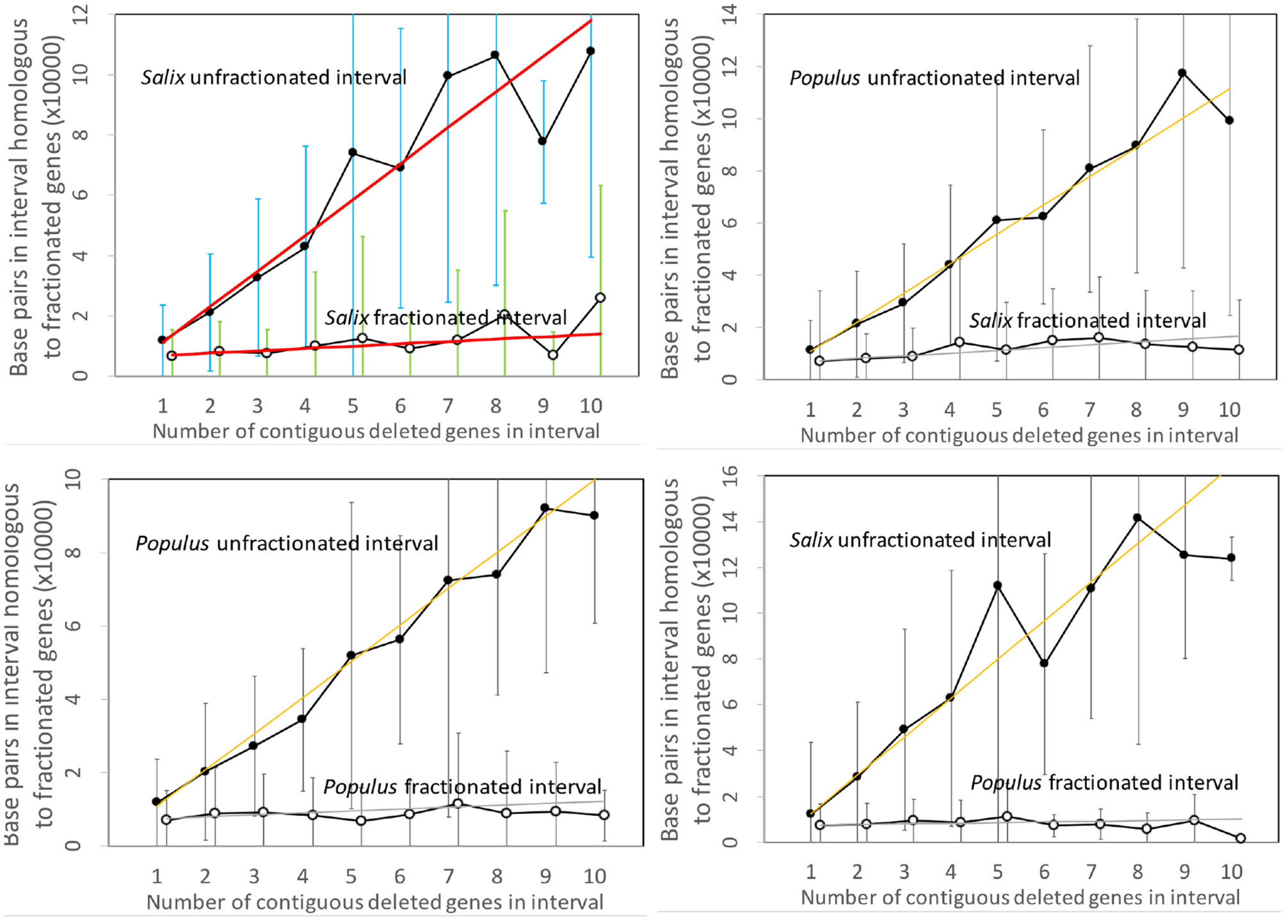


FIGURE 3 | Comparison of DNA content in unfractionated and fractionated intervals in the *Salix* and *Populus* genomes. Linear regression fits are indicated. Self-comparisons on the left do not distinguish between subgenomes since these are hard to identify across chromosomes and are generally mingled due to interchromosomal rearrangements, such as reciprocal translocation and chromosome fission and fusion. The two comparisons between genomes on the right hand side analyze gene loss from each genome separately. We use the terms "fractionated" and "unfractionated" in these two panels to mean "reduced" and "conserved," even though the polyploidization-induced fractionation does not play a role here.

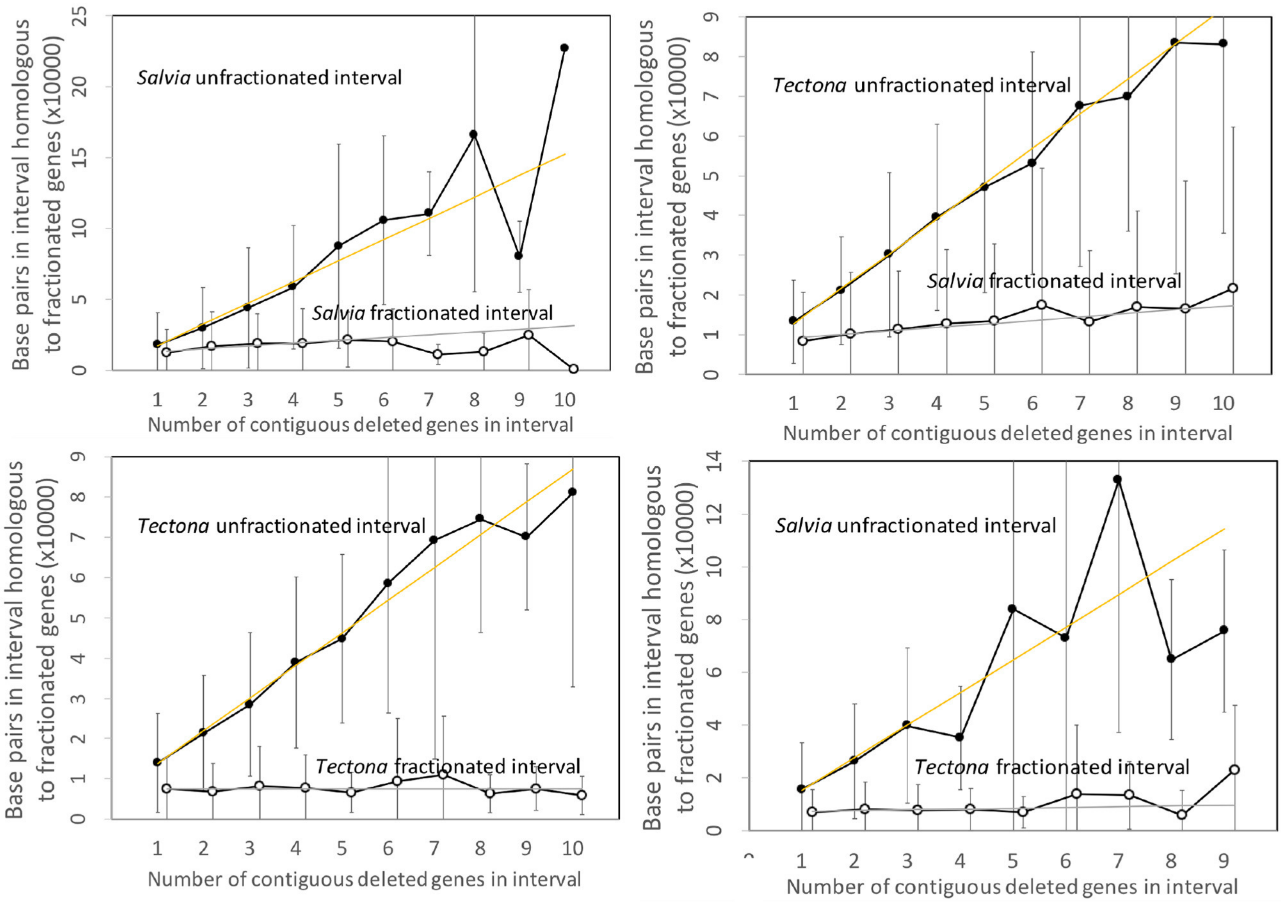


FIGURE 4 | Comparison of DNA content in unfractionated and fractionated intervals in the *Salvia* and *Tectona* genomes.

of the gene density for that interval. For the fractionated, or totally reduced, side, the number of base pairs per missing gene provides an upper limit (via its inverse) on the number of full-length pseudogenes that may be in the interval. Although most pseudogene tools were developed in the context of human or vertebrate genomes, and have limited applicability for plant genomes (Xiao et al., 2016), Xie et al. have succeeded in implementing PSEUDOPIPE (Zhang et al., 2006) for surveying pseudogenes in a range of plant species, and their results will be seen to be consistent with ours in the analyses below.

3. RESULTS

3.1. Willow and Poplar

Figure 3 contains the results of our analysis of the *Salix* and *Populus* genomes. The two panels on the left show the expected approximate linear growth in the number of base pairs in the unfractionated side of the interval. The great variability of the individual regions simply reflects the inhomogeneity of gene density along the length of the chromosome. In contrast, the regions in both *Salix* and *Populus* that have lost annotated genes show zero growth, with relatively little variability, as a function of the number of missing genes; they have lost almost all their DNA sequence. There cannot be significant numbers of pseudogenes, full or reduced, or other relics of the missing genes. This is striking evidence in favor of the predominance of excision.

3.2. *Salvia* and Teak

Figure 4 contains the results of the corresponding analysis of the *Salvia* and *Tectona* genomes. The figures are very similar to those from the Salicacea. Some of the curves show great fluctuation of the values for the longer intervals, but this is likely due to smaller sample size. Of interest is that the DNA content of the fractionated (read: "reduced") intervals formed after speciation show a small but steady increase, but still orders of magnitude less than the sizes of the unfractionated ("conserved") intervals.

3.3. Flax and Rubber

Figure 5 repeats the same analysis, this time applied to the *Linum* and *Hevea* genomes. The results parallel those of the two other pairs of genomes, except for the apparently anomalous behavior of the *Hevea* intervals, where the number of base pairs attains the same level as the conserved genes in *Linum*. This, however, may be seen as an artifact of the disproportionately large genome

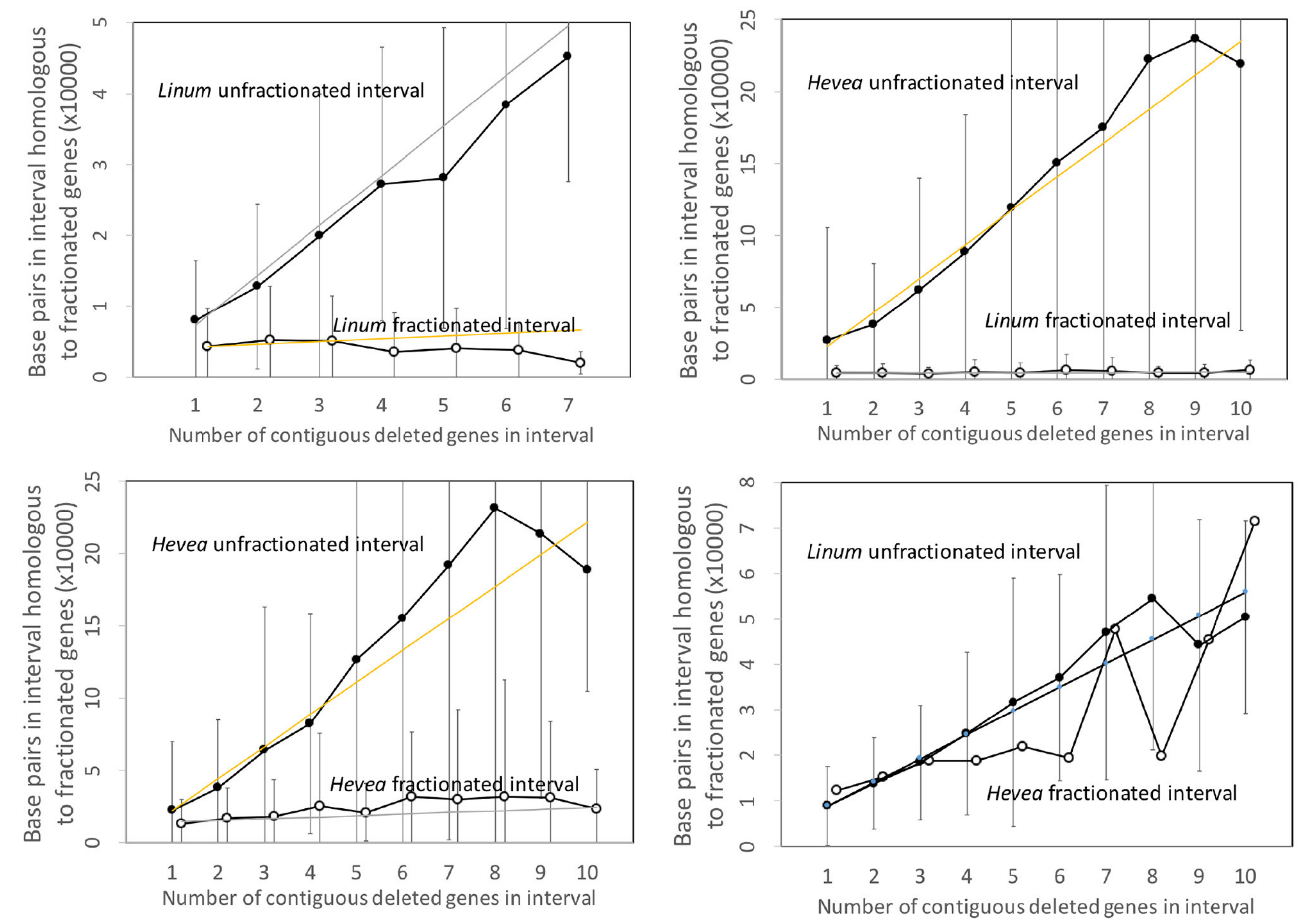


FIGURE 5 | Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Linum* and *Hevea* genomes.

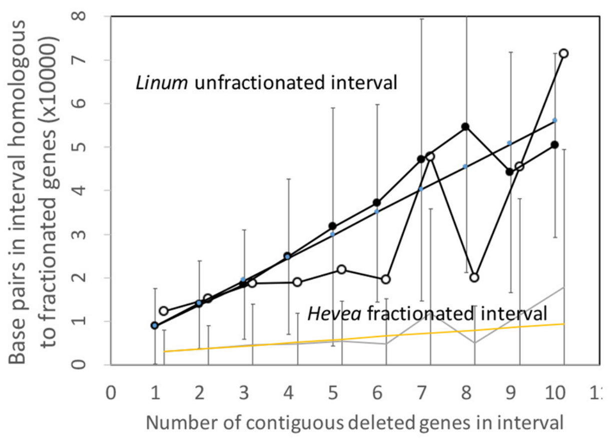


FIGURE 6 | Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Linum* and normalized *Hevea* genomes.

of *Hevea* with respect to that of *Linum*. The intergenic space in *Hevea* is four or five times as great as that of *Linum*, and there is much scope for retention or acquisition of repetitive elements and other sequence over the long period since the speciation event, which occurred much earlier than the other events we study.

To put this disproportions in perspective, we can normalize the *Hevea* results by a factor which measures the difference in sizes of the two genomes. This produces the comparisons in **Figure 6**, which better resembles those of the Salicaceae and Lamiaceae.

3.4. Pear and Apple

Figure 7 shows the analysis of the *Pyrus* and *Malus* genomes. Here again, we have an anomalous large amount of DNA in the *Malus* reduced gene intervals after speciation. It is true that the *Malus* genome is larger than *Pyrus*, but explaining this through normalization (**Figure 8**) is not completely satisfactory. This is the only trend out of the thirty-two we have presented that departs from our main narrative.

3.5. Comparisons Across Genome Pairs

To compare the results from the four pairs of genomes, we must take into account the diverse genome sizes, number of genes in a genome, and the resulting gene densities. **Figure 9** shows

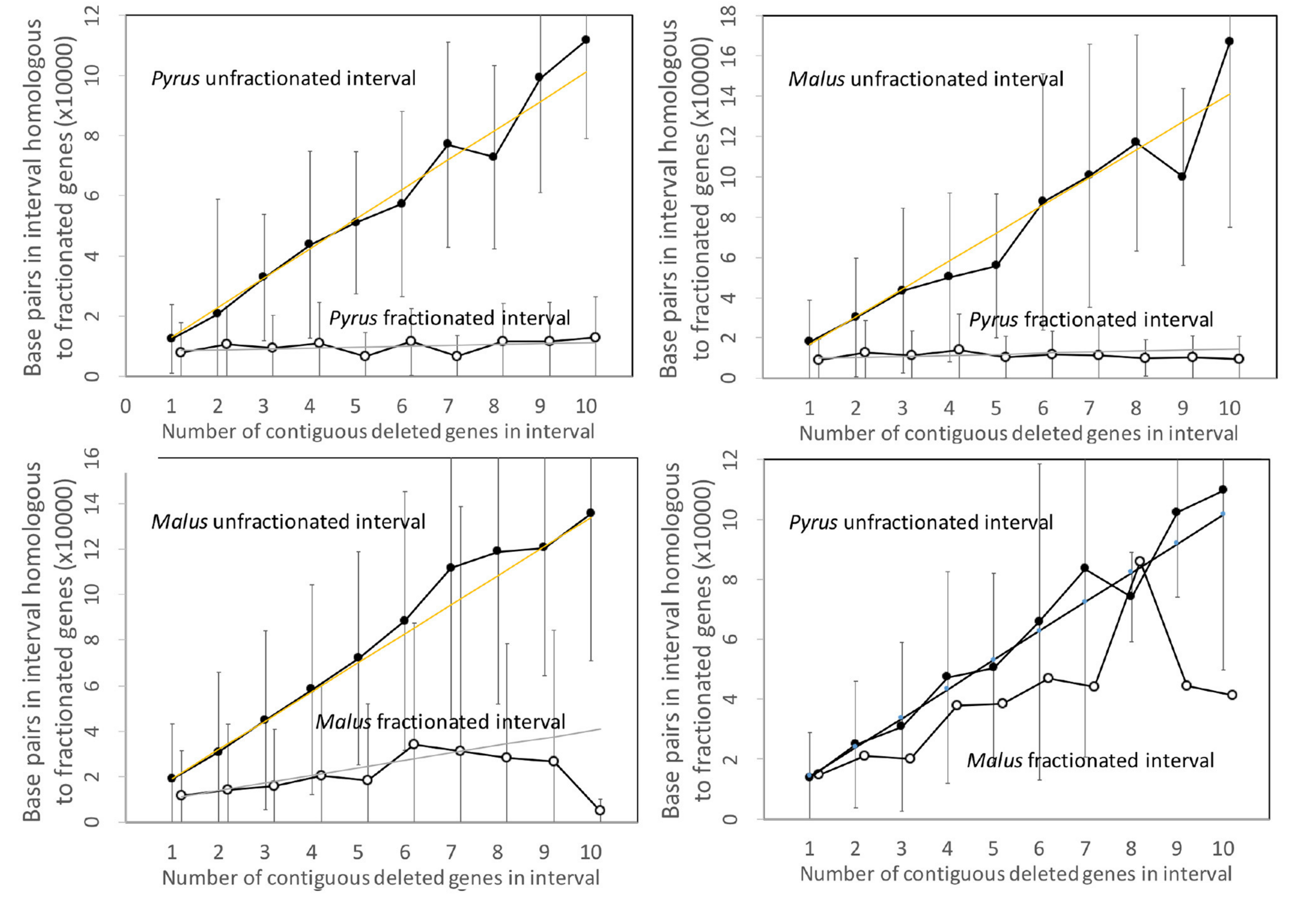


FIGURE 7 | Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Pyrus* and *Malus* genomes.

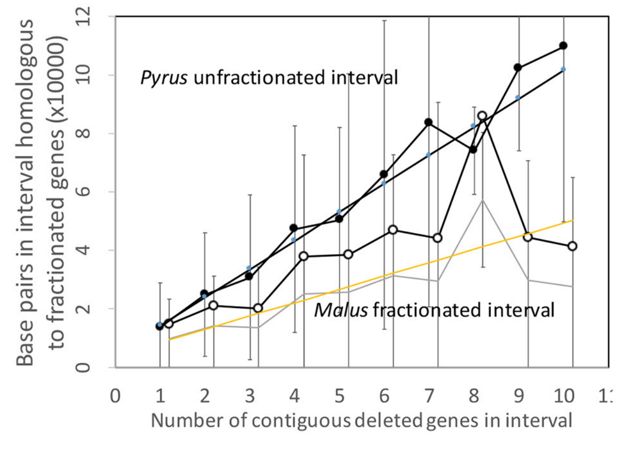


FIGURE 8 | Comparison of DNA content in conserved and reduced intervals in the *Pyrus* and normalized *Malus* genomes.

that gene density (or rather its inverse: base pairs per length of conserved fragment) in unfractionated and conserved intervals closely tracks the average gene density (or its inverse) for the entire genome. At the same time, the residual sequence length in intervals where fractionation or gene loss has taken place is not sensitive to gene density, it remains very close to zero, as expected from an excision explanation.

We can also report, although it seems superfluous after examining **Figures 3–7**, **9**, that a *t*-test confirms at a very high level of significance that the slopes of the two regressions in each panel are different.

3.6. Occurrences of Gene Translocation

To exclude other explanations of our syntenic block data, such as that in **Figure 1E**, we looked further into the fate of the fractionated genes in the *Populus-Salix* comparison. By setting the minimum block size to 1 in the SynMap self-comparison, we could detect all pairs of gene duplicates, not only those in synteny blocks. We then searched for pairs to the singletons identified in the original (default 5) construction of synteny blocks that we analyzed in section 3.1 above. Of the 429 out of 8,307 *Salix* singletons, we found only 429, or 5%, that were paired else where in the genome at approximately the expected similarity level. Of the 10,737 *Populus* singletons, only 742, or 7%, were paired elsewhere. Moreover, some of the

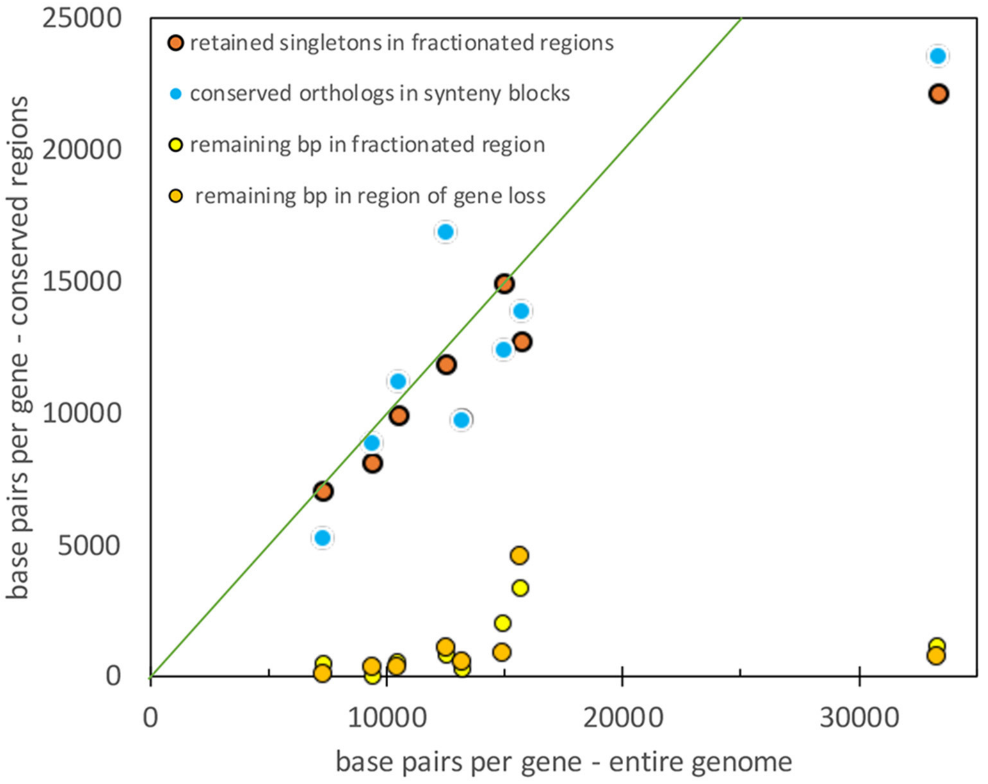


FIGURE 9 | Comparison of gene density in unfractionated regions and the whole genome. Diagonal represents equality between the two densities.

pairs that were identified could have been distinct paralogs that were part of a pre-existing triplet before fractionation—such triplets or higher sets of paralogs are not uncommon. We can conclude that translocation as an alternative explanation to excision can account for only a very small fraction of the gaps in synteny blocks.

There remains the possibility that if the missing genes did not translocate out of the synteny block, the singletons may have migrated in, after the polyploidization or speciation event (Vicient and Casacuberta, 2017). The main mechanism for this would be retrotransposition. However, retroposons are generally not annotated as genes in the CoGe database, even in the unmasked genome sequences we studied, and thus would not show up as singletons. Neither are many of the singletons likely to be translocated genes: a large proportion of genes in these genomes are paired, and an equal proportion of the putatively translocated singletons would show up as pairs elsewhere in the genome in the minimum block size 1 analysis. We have already seen that this is not the case.

4. CONCLUSIONS

The statistical evaluation of the massive duplicate gene cohorts created by speciation or polyploidization shows that pseudogenization is either a very rare process or does not result in much stable structure. By the present time, the clear impression is that fractionation simply excises the DNA of a gene or several contiguous genes. Ongoing work to be reported elsewhere suggests that this elimination of sequence does occur piecemeal over 30 million years or even 1 million years. It is of course still possible that once a pseudogene is created, or a gene otherwise silenced, its DNA is immediately vulnerable to repeated small deletions, so that the pseudogene itself would be transient. The distinction between this and some single-event excision becomes a matter of semantics.

More surprising perhaps is that gene loss after speciation, occurring independently in two sister genomes, seems to follow the same trajectory. There is of course no genomic interaction between species pairs like *Salvia* and *Tectona*, but their common origin allows us to use one to track the gene loss pattern in the other. There remain questions of how universal excision is; in the *Salvia-Tectona* and *Poplar-Salix* comparisons it is very clear. Because of the genome size differential, it is harder to determine in *Linum-Hevea*, while in the case of *Malus*, though fractionation proceeds by excision, further gene loss may involve other mechanisms as well. We note that the role of differential amounts of repetitive sequence and active retroposon activity can impact this type of comparison between species, less so within one species.

Although it is difficult to say if it has any impact on our analysis, we note that speciation of apple and pear came later than their common whole genome duplication. It is the same for poplar and willow. The teak whole genome duplication occurred before speciation, but the *salvia* came after. That means that we analyzed more recent *salvia* fractionation than an earlier one that it shares with teak. The rubber-flax speciation is much more ancient than their individual whole genome duplications.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://genomevolution.org/coge/.

AUTHOR CONTRIBUTIONS

ZY and DS planned the research, carried it out, and wrote this article. CZ developed and organized the data and participated in the planning and devising the analyses. VA contributed to the interpretation of the analyses and to understanding the pertinence of the results. All authors contributed to the article and approved the submitted version.

FUNDING

Research and publication costs were supported in part by Discovery Grant number 8867-200 from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

REFERENCES

Byrnes, J. K., Morris, G. P., and Li, W. H. (2006). Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol. Biol. Evol.* 23, 1136–1143. doi: 10.1093/molbev/msj121

Eckardt, N. A. (2001). A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell* 13, 1699–1704. doi: 10.1105/tpc.13.8.1699

Hansen, M. T. (1978). Multiplicity of genome equivalents in the radiation-resistant bacterium *Micrococcus radiodurans*. *J. Bacteriol.* 134, 71–75. doi: 10.1128/JB.134.1.71-75.1978

Jacq, C., Miller, J. R., and Brownlee, G. G. (1977). A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* 12, 109–120. doi: 10.1016/0092-8674(77)90189-1

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673. doi: 10.1111/j.1365-313X.2007.03326.x

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., et al. (2008). Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiol.* 148, 1772–1781. doi: 10.1104/pp.108.124867

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer.

Sankoff, D., Zheng, C., Wang, B., and Fernando Buen Abad Najar, C. (2015). Structural *vs.* functional mechanisms of duplicate gene loss following whole genome doubling. *BMC Genomics* 15:915. doi: 10.1109/ICCABS.2014.6863915

Sankoff, D., Zheng, C., and Zhu, Q. (2010). The collapse of gene complement following whole genome duplication. *BMC Genomics* 11:313. doi: 10.1186/1471-2164-11-313

Tobiason, D. M., and Seifert, H. S. (2006). The obligate human pathogen, Neisseria gonorrhoeae, is polyploid. *PLoS Biol.* 4:e185. doi: 10.1371/journal.pbio.0040185

van Hoek, M. J., and Hogeweg, P. (2007). The role of mutational dynamics in genome shrinkage. *Mol. Biol. Evol.* 24, 2485–2494. doi: 10.1093/molbev/msm183

Vicient, C. M. and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann Bot.* 120, 195–207. doi: 10.1093/aob/mcx078

Winge, Ö. (1917). The chromosomes: their number and general importance. *Comptes Rendus des Travaux Lab. Carlsberg* 13, 131–275.

Wolfe, K., and Shields, D. (1977). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.

Xiao, J., Sekhwal, M. K., Li, P., Ragupathy, R., Cloutier, S., Wang, X., et al. (2016). Pseudogenes and their genome-wide prediction in plants. *Int. J. Mol. Sci.* 17, 1991–2006. doi: 10.3390/ijms17121991

Xie, J., Li, Y., Liu, X., Zhao, Y., Li, B., Ingvarsson, P. K., et al. (2019) Evolutionary origins of pseudogenes and their association with regulatory sequences in plants. *Plant Cell* 31, 563–578. doi: 10.1105/tpc.18.00601

Yu, Z., Zheng, C., and Sankoff, D. (2020). Gaps and runs in syntenic alignments, in *International Conference on Algorithms for Computational Biology*. Lecture Notes in Computer Science 12099, 49–60.

Yu, Z. N., and Sankoff, D. (2016). A continuous analog of run length distributions reflecting accumulated fractionation events. *BMC Bioinformatics* 17(Suppl. 14):412. doi: 10.1186/s12859-016-1265-5

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* 1437–1439. doi: 10.1093/bioinformatics/btl116

Zheng, C., Wall, P. K., Leebens-Mack, J., dePamphilis, C., Albert, V. A., and Sankoff, D. (2009). Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* diploid. *J. Bioinform. Comput. Biol.* 7, 499–520. doi: 10.1142/s0219720009004199

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.