POLYPLOID POPULATION GENETICS AND EVOLUTION - FROM THEORY TO PRACTICE

EDITED BY: Hans D. Daetwyler and Richard John Abbott

PUBLISHED IN: Frontiers in Ecology and Evolution, Frontiers in Genetics and Frontiers in Plant Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-390-6
DOI 10.3389/978-2-88963-390-6

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: researchtopics@frontiersin.org

POLYPLOID POPULATION GENETICS AND EVOLUTION - FROM THEORY TO PRACTICE

Topic Editors:
Hans D. Daetwyler, La Trobe University, Australia
Richard John Abbott, University of St Andrews, United Kingdom

Citation: Daetwyler, H. D., Abbott, R. J., eds. (2020). Polyploid Population Geneics and Evolution - From Theory to Practice. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-390-6

Table of Contents

frontiers in Ecology and Evolution

EDITORIAL
published: 28 November 2019
doi: 10.3389/fevo.2019.00460

Editorial: Polyploid Population Genetics and Evolution—From Theory to Practice

Abdulqader Jighly[1]*, *Richard J. Abbott*[2] *and Hans D. Daetwyler*[1,3]

[1] Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, VIC, Australia, [2] School of Biology, University of St Andrews, St Andrews, United Kingdom, [3] School of Applied Systems Biology, La Trobe University, Melbourne, VIC, Australia

Keywords: allopolyploids, autopolyploids, polyploidization, mathematical biology, simulation

OPEN ACCESS

Approved by:
Genlou Sun,
Saint Mary's University, Canada

Correspondence:
Abdulqader Jighly
abdulqader.jighly@agriculture.vic.gov.au

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

Received: *06 September 2019*
Accepted: *15 November 2019*
Published: *28 November 2019*

Citation:
Jighly A, Abbott RJ and Daetwyler HD (2019) Editorial: Polyploid Population Genetics and Evolution—From Theory to Practice. Front. Ecol. Evol. 7:460. doi: 10.3389/fevo.2019.00460

Editorial on the Research Topic

Polyploid Population Genetics and Evolution—From Theory to Practice

Despite polyploids being widespread and of great importance in eukaryotic diversification, our understanding of the dynamics of the evolution and inheritance of polyploids is less advanced than for diploids. The challenges in studying the population genetics and evolution of polyploids reside in the presence of more than two homoeologous "diverged but related" chromosome copies in allopolyploids or homologous "identical" chromosome copies in autopolyploids. Moreover, diploidization processes following polyploidy trigger other challenges in inferring paleo-polyploidization or ancient polyploidization events, which complicate the study of diverged homo(eo)logous genes and modeling of ecological factors affecting polyploids and their interactions with diploid ancestors. Statistical methods originally developed for the diploid mode of inheritance are generally biased when analyzing polyploids creating an urgent need to develop new methods for studying the evolutionary dynamics and modes of inheritance of polyploids (Dufresne et al., 2014; Meirmans et al., 2018). The aim of this Research Topic is to enhance our current understanding of the population genetics and evolution of polyploids and to highlight the practical applications that flow from such understanding. The collection of 12 papers covers four main areas of investigation: (1) the establishment of polyploids and long-term evolutionary consequences of polyploidy; (2) the evolution of gene expression, gene families, and chromosomes in polyploids; (3) the development of novel statistical polyploid-friendly population genetics models; and (4) the practical applications of different statistical models in polyploid trait evolution, quantitative genetics, and plant breeding.

With regard to the first of these topics, Baduel et al. provide a comprehensive review of factors affecting the successful establishment of newly formed polyploids in the wild and the short- and long-term costs and benefits that emanate from polyploidy. In this context, they discuss recent relevant ecological, physiological, cytological and genomic research, and underline the "*wondrous cycles*" of polyploidy (Wendel, 2015) in which polyploidization repeatedly happens after diploidization events over long evolutionary timescales. The advantages and disadvantages of polyploidy are further considered by Gaynor et al., but from the standpoint of a macro-scale study of the effects of polyploidization on the geographical community structure of two widely distributed flowering plant families, the Brassicaceae and Rosaceae, both of which have experienced multiple rounds of polyploidization events in the past. By combining cytogeographical information with phylogenetic analyses of plant communities in these two families across the USA, they show that communities may be shaped in diverse ways by polyploidy, but that impacts of genome duplication are not clear cut and are lineage specific. They highlight the need for much greater information on ploidal variation across species' ranges to provide a deeper understanding of the effects of genome duplication on plant community structure.

Following polyploidization, alterations to chromosome number and structure as well as gene function, expression, and copy number may occur and feature prominently in diploidization (Ohno, 1970; Tate et al., 2009; Conant et al., 2014; Jighly et al., 2019). With regard to changes in chromosome number, Jelenić et al. develop a mitotic mathematical model to predict the chromosome loss rate in polyploids before testing it in polyploid cells of the yeast, *Saccharomyces cerevisiae*. The model depends on spindle dynamics and the maximum duration of mitotic arrest. They show that a small change in spindle assembly time can cause a massive increase in the rate of chromosome loss in tetraploid cells. Focusing on gene expression and function, Takahagi et al. analyze 727 previously published RNA sequence datasets of hexaploid wheat collected from different developmental stages, tissues, and environmental conditions to examine differences in expression profiles. They observe genes that are present and expressed in triplets, doublets, or specifically in one subgenome, contributing to broad biological functions and annotations. With regard to gene family changes, Mable et al. report an analysis of European diploid and tetraploid *Arabidopsis lyrata* and *Arabidopsis arenosa* populations to infer the complex evolution of the "*S-receptor kinases*" (SRK) gene family. This gene family is involved in the female component of genetically controlled self-incompatibility and is subject to strong balancing selection (Castric and Vekemans, 2007). In turn, they examine how the diversity of *SRK* alleles in tetraploids compares with that in diploid relatives, whether there is increased trans-specific polymorphism in tetraploids for these genes, if introgression occurs among species and ploidy levels, and whether copy number variation exists among paralogs.

Developing and extending widely used diploid theories and statistical models to fit polyploids is an important aim of the Research Topic. Meirmans and Liu extend the widely used analysis of molecular variance (AMOVA) to autopolyploids. This can be regarded as a significant step forward, given that since AMOVA was first developed by Excoffier et al. (1992), it has been widely employed in analyzing the population genetics of diploids. Similarly, site frequency spectrum (SFS) based methods such as the neutrality (Tajima, 1989) and (Fay and Wu, 2000) as well as heterozygosity of allelic variant tests such as Tajima's estimator of nucleotide diversity (Tajima, 1983) are widely used in diploid population genetics. Together with other SFS methods applied to high-throughput sequencing data, Ferretti et al. extend their application to autopolyploid populations and discuss their bias when applied to small populations. Detecting gene copy number variation is one of the most challenging tasks in the population genetic analysis of autopolyploids, leading (Knaus and Grünwald) to develop an R package "*VCFR*" to infer copy number variation in polyploids. The novelty of their method is that it does not require including the copy number of genomic regions (or alleles) *a priori*, but instead, VCFR infers them depending on the frequency of the most abundant alleles. Bourke et al. review the existing methods applied to experimental autopolyploid populations, such as breeding populations. They focus on methods of genotyping of polyploids, physical and genetic mapping procedures, simulating polyploid breeding populations, and quantitative genetic analyses including quantitative trait loci (QTL) mapping, genome wide association studies (GWAS) and genomic prediction in polyploids.

The final papers that comprise this Research Topic focus on the applications of polyploid population genetics in plant breeding. Ferrão et al. use a large breeding population of 1,575 autotetraploid blueberry individuals to dissect the genetic basis of eight fruit related traits and detect QTL associated with genotyping-by-sequencing based single nucleotide polymorphism (SNP) markers. They call their SNPs twice, with diploid and tetraploid genotype coding to compare the effect of diploid-like calling on GWAS results. Diploid coding resulted in shorter linkage disequilibrium blocks and a much smaller number of significantly associated QTL indicating the importance of using a tetraploid model. As an alternative to using tetraploid SNP coding, Manrique-Carpintero et al. developed a dihaploid potato population and conducted QTL mapping for vigor, height, and different tuber traits. Finally, in an examination of a population of synthetic allohexaploid wheat (*Triticum turgidum* – AABB × *Aegilops tauschii* – DD), Jighly et al. divided the additive variance for 12 biotic and abiotic stresses among the 21 chromosomes representing the A, B, and D subgenomes. They found that the wild D subgenome had the highest contribution to the additive variance in most traits, while the A subgenome had the lowest. They also reported a weak but significant positive correlation between the cumulative size of each of three homoeologous chromosomes and their cumulative additive variance.

The articles published on this Research Topic provide a body of knowledge in the field of polyploid population genetics and evolution. Though much progress has been made in this area, many challenges remain. Of particular importance will be the further development of robust statistical models for polyploids and the effective and efficient simulation of their population genetic and genomic complexities (Dufresne et al., 2014; Jighly et al., 2018). This will allow the testing of models and assumptions under complex evolutionary and demographic scenarios with validation using empirical data. Given the economic, ecological, and evolutionary importance of polyploidy, further concentrated research efforts are required to advance population genetic theories and applications that relate directly to polyploid species.

AUTHOR CONTRIBUTIONS

AJ provided the first draft. RA and HD edited the manuscript and provided additional text. All authors revised and approved the manuscript for submission.

REFERENCES

Castric, V., and Vekemans, X. (2007). Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? *BMC Evol. Biol.* 7:132. doi: 10.1186/1471-2148-7-132.

Conant, G. C., Birchler, J. A., and Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin Plant Biol.* 19, 91–98. doi: 10.1016/j.pbi.2014.05.008

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581.

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial-DNA restriction data. *Genetics* 131, 479–491.

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive darwinian selection. *Genetics* 15, 1405–1413. Available online at: http://www.genetics.org/content/155/3/1405.long

Jighly, A., Joukhadar, R., Sehgal, D., Singh, S., Ogbonnaya, F. C., and Daetwyler, H. D. (2019). Population-dependent reproducible deviation from natural bread wheat genome in synthetic hexaploid wheat. *Plant J.* 100, 801–812. doi: 10.1111/tpj.14480

Jighly, A., Lin, Z., Forster, J. W., Spangenberg, G. C., Hayes, B. J., and Daetwyler, H. D. (2018). Insights into population genetics and evolution of polyploids and their ancestors. *Mol. Ecol. Res.* 18, 1157–1172. doi: 10.1111/1755-0998.12896

Meirmans, P. G., Liu, S., and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Hered.* 109, 283–296. doi: 10.1093/jhered/esy006

Ohno, S. (1970). *Evolution by Gene Duplication*. New York, NY: Springer-Verlag.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Tate, J. A., Joshi, P., Soltis, K. A., Soltis, P. S., and Soltis, D. E. (2009). On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biol.* 9:80. doi: 10.1186/1471-2229-9-80

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102, 1753–1756. doi: 10.3732/ajb.1500320

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 Ecology and Evolution

REVIEW
published: 20 August 2018
doi: 10.3389/fevo.2018.00117

The "Polyploid Hop": Shifting Challenges and Opportunities Over the Evolutionary Lifespan of Genome Duplications

Pierre Baduel[1], *Sian Bray*[2], *Mario Vallejo-Marin*[3], *Filip Kolář*[4,5] *and Levi Yant*[2,6]*

[1] Institut de Biologie de l'École Normale Supérieure, Paris, France, [2] Department of Cell and Developmental Biology, John Innes Centre, Norwich, United Kingdom, [3] Biological and Environmental Sciences, University of Stirling, Stirling, United Kingdom, [4] Department of Botany, University of Innsbruck, Innsbruck, Austria, [5] Department of Botany, Faculty of Science, Charles University in Prague, Prague, Czechia, [6] School of Life Sciences and Future Food Beacon, University of Nottingham, Nottingham, United Kingdom

OPEN ACCESS

Edited by:
*Richard John Abbott,
University of St Andrews,
United Kingdom*

Reviewed by:
*Barbara K. Mable,
University of Glasgow,
United Kingdom
Christian Parisod,
University of Neuchâtel, Switzerland
Andreas Madlung,
University of Puget Sound,
United States*

***Correspondence:**
*Levi Yant
leviyant@gmail.com*

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

Received: 25 April 2018
Accepted: 23 July 2018
Published: 20 August 2018

Citation:
*Baduel P, Bray S, Vallejo-Marin M,
Kolář F and Yant L (2018) The
"Polyploid Hop": Shifting Challenges
and Opportunities Over the
Evolutionary Lifespan of Genome
Duplications. Front. Ecol. Evol. 6:117.
doi: 10.3389/fevo.2018.00117*

The duplication of an entire genome is no small affair. Whole genome duplication (WGD) is a dramatic mutation with long-lasting effects, yet it occurs repeatedly in all eukaryotic kingdoms. Plants are particularly rich in documented WGDs, with recent and ancient polyploidization events in all major extant lineages. However, challenges immediately following WGD, such as the maintenance of stable chromosome segregation or detrimental ecological interactions with diploid progenitors, commonly do not permit establishment of nascent polyploids. Despite these immediate issues some lineages nevertheless persist and thrive. In fact, ecological modeling commonly supports patterns of adaptive niche differentiation in polyploids, with young polyploids often invading new niches and leaving their diploid progenitors behind. In line with these observations of polyploid evolutionary success, recent work documents instant physiological consequences of WGD associated with increased dehydration stress tolerance in first-generation autotetraploids. Furthermore, population genetic theory predicts both short- and long-term benefits of polyploidy and new empirical data suggests that established polyploids may act as "sponges" accumulating adaptive allelic diversity. In addition to their increased genetic variability, introgression with other tetraploid lineages, diploid progenitors, or even other species, further increases the available pool of genetic variants to polyploids. Despite this, the evolutionary advantages of polyploidy are still questioned, and the debate over the idea of polyploidy as an evolutionary dead-end carries on. Here we broadly synthesize the newest empirical data moving this debate forward. Altogether, evidence suggests that if early barriers are overcome, WGD can offer instantaneous fitness advantages opening the way to a transformed fitness landscape by sampling a higher diversity of alleles, including some already preadapted to their local environment. This occurs in the context of intragenomic, population genomic, and physiological modifications that can, on occasion, offer an evolutionary edge. Yet in the long run, early advantages can turn into long-term hindrances, and without ecological drivers such as novel ecological niche availability or agricultural propagation, a restabilization of the genome via diploidization will begin the cycle anew.

Keywords: polyploidy, selection, population genetics, evolution, autopolyploidy, genome duplication

INTRODUCTION

Whole genome duplication (WGD) is a pervasive event in the evolution of eukaryotes, with an especially strong representation throughout the plant kingdom. Despite this prevalence, however, WGD is no easy victory: the presence of an extra set of chromosomes creates numerous biological challenges ranging from chromosome mis-segregation to altered gene expression, changes in cell size or in intracellular physiology (Ramsey and Schemske, 2002; Osborn et al., 2003; Adams and Wendel, 2005; Comai, 2005; Chen and Ni, 2006; Chen, 2007; Otto, 2007; Parisod et al., 2010; Brownfield and Köhler, 2011; Hollister, 2015). Numerous studies have shown evidence of dysfunction in newly formed polyploids across kingdoms: in plants, fungi and animals, including notably cancer cells (Ramsey and Schemske, 2002; Storchova and Pellman, 2004; Comai, 2005; Yant and Bomblies, 2017). Even more, polyploid establishment is substantially constrained by overall low chances of polyploid mutants persisting among their diploid progenitors (Levin, 2002; Ramsey and Schemske, 2002). This stems from both direct competition between the two cytotypes (Yamauchi et al., 2004) and frequency dependent selection (Levin, 1975), which suggests most autopolyploids are likely to go extinct before establishment (Levin, 1975; Husband, 2000).

Despite these challenges, WGD events have occurred repeatedly throughout the evolution of eukaryotes (Gregory and Mable, 2005; Wood et al., 2009; Wendel, 2015), leading to an abundance of established polyploid species in the wild. There is clear evidence of WGD in the ancestry of most plant lineages (Vision et al., 2000; Bowers et al., 2003; Paterson et al., 2004; Schlueter et al., 2004; Pfeil et al., 2005; Barker et al., 2008; Burleigh et al., 2008; Shi et al., 2010). Among angiosperms, it is estimated that 30–70% have undergone additional WGD events (Stebbins, 1938; Grant, 1963; Goldblatt, 1980; Masterson, 1994; Wood et al., 2009; Mayrose et al., 2011; Ruprecht et al., 2017). Polyploidy is also common among important crops (e.g., wheat, maize, sugar cane, coffee, cotton, potato, and tobacco), suggesting that WGD is often a key factor in successful crop domestication (Salman-Minkov et al., 2016). Obtaining precise estimates of the extent of polyploidy can be complicated, in part due to difficulties obtaining direct empirical evidence. For example, autopolyploids, resulting from within-species genome duplication, are often not considered a separate species from their diploid progenitors. As a result, their overall abundance compared to allopolyploids (where two distinct genomes are combined) have historically been underestimated (Soltis et al., 2007; Barker et al., 2016).

Polyploidy, a High-Risk, High-Gain Path

Frequency estimates of WGD increase in habitats affected by environmental disturbances (Favarger, 1984; Brochmann et al., 2004; Parisod et al., 2010). Concordant with this observation, in diploid-polyploid systems overlapping former glaciation limits, polyploids are found more frequently in the previously glaciated areas while their diploid progenitors commonly remain or retreat within former refugia (Ehrendorfer, 1980; Kadereit, 2015). For example, in *Arabidopsis,* both auto and allopolyploidization events were estimated to coincide with glacial maxima (Novikova et al., 2018). Beyond the fact that environmental stochasticity both increases the rate of WGD and provides new space for colonization (Baack, 2005; Fawcett et al., 2009; Oswald and Nuismer, 2011), such observations implicate WGD in speciation and adaptive radiation (Wood et al., 2009) and support the long-standing hypothesis that WGD *per se* can potentiate evolutionary adaptation, although evidence for this is somewhat mixed. Clear empirical evidence from *in vitro* evolution experiments in yeast demonstrated that tetraploids adapted faster than lower ploidies (Selmecki et al., 2015) and has bolstered this hypothesis. However, complementary approaches such as ecological niche modeling do not always support niche innovation in polyploids (Glennon et al., 2014). For example in primroses, the niches occupied by the three polyploid species (tetraploid, hexaploidy and octoploid) were distinct relative to the diploid progenitor but they were also narrower (Theodoridis et al., 2013).

Here we synthesize recent advances in polyploid research from new genomic, ecological, and cytological analyses with older observations and theoretical arguments into two primary dimensions (**Figure 1**): consequences (challenges vs. gains) of WGD and their time-span (short-term vs. long-term). To address specifically the effects of WGD, we focus on autopolyploids, which arise from within-species WGD events and thus carry four or more homologous copies of each chromosome (for a clear depiction, see Bomblies and Madlung, 2014). We thus largely set aside allopolyploids (polyploid hybrids), which strongly confound the effects of WGD with hybridization. On many fronts, recent results from autopolyploid systems have confirmed earlier theoretical predictions, but some have unveiled surprising new results in the context of a wide range of biological processes. Most strikingly, the population genetic consequences of WGD have been the subject of ample theoretical arguments despite thin experimental support to date, but this is changing. Our synthesis paints an overall picture of autopolyploidy as a high-risk high-gain path, where long-term complications often outweigh initial benefits while paving the way for re-diploidization. This depiction strengthens the idea of polyploidy as a transitionary state, a "hop," which has seen growing support from the polyploid community (Escudero et al., 2014; Wendel, 2015).

THE SHORT-TERM CHALLENGES

Meiosis

Perhaps the most stringent challenge faced by a nascent autopolyploid is directly tied to the very process of reproduction. A sudden doubling of homologous chromosome number disrupts regular meiotic pairing and segregation: instead of each chromosome having only one homolog with which to pair, in autotetraploids there are suddenly three.

The situation in most diploids is relatively straightforward, with proper chromosome pairing during synapsis typically relying on programed double-strand DNA breaks and a sequence-based homology search for the homolog using these broken fragments (Grelon et al., 2001; Page and Hawley, 2003; Stacey et al., 2006; Hartung et al., 2007). Once homologous

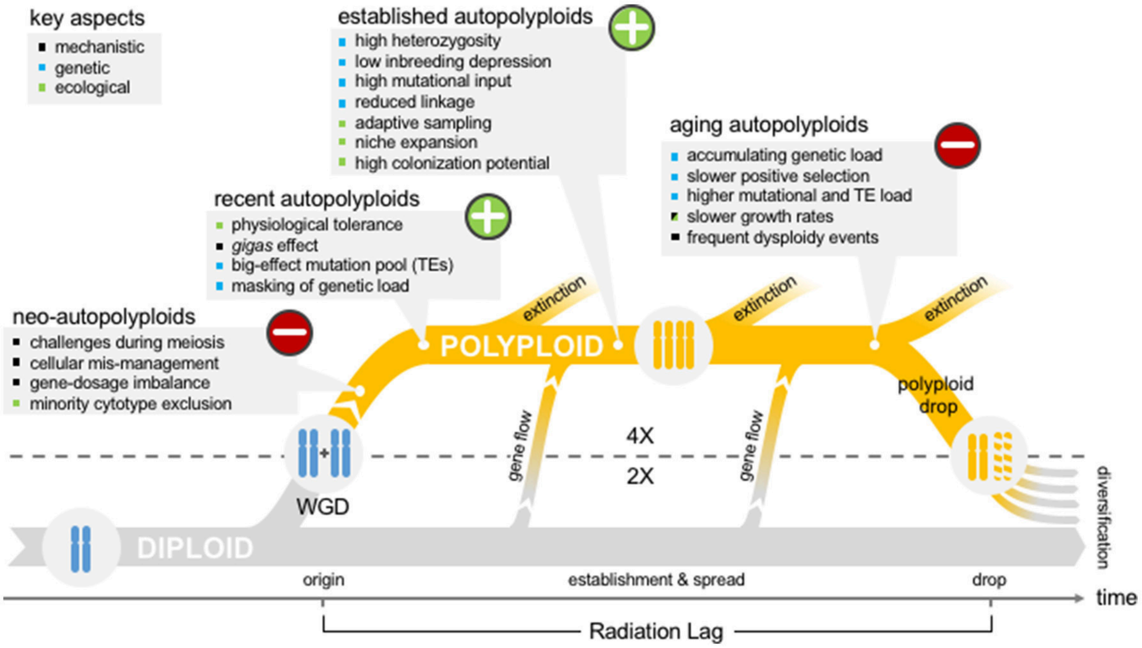


FIGURE 1 | The polyploid hop. Schematic representation of the temporal evolution of a diploid-autopolyploid system accompanied with the shifting benefits and pitfalls linked with WGD and that can contribute to the different stages of polyploid evolution.

chromosomes have aligned, a small fraction of these breaks mature into crossovers (COs) between homologes, thus creating bivalent chromosome pairs physically linked by the CO. These bivalents then align parallel to the poles at which point the COs are essential for the creation of mechanical tension between the assembling spindles via connections to each centromere. This tension transmitted through the obligate COs ensures the correct orderly segregation of chromosomes and further acts as an essential cell cycle checkpoint allowing progression to anaphase (Lampson and Cheeseman, 2011; Campbell and Desai, 2013).

In newly formed polyploids however, the presence of multiple equivalent partners for each chromosome leads to more complex arrays of CO formations. If there is more than one crossover per bivalent, a single homolog can have two separate partners, resulting in a multivalent. Most multivalent configurations are not conducive to the formation of regular tension (Bomblies et al., 2016) and some entirely fail to involve one homolog, leading to mis-segregation and aneuploidy. Compounding this, in nascent autopolyploids entanglements and interlocks can occur between non-homologs. If left unresolved these can result in catastrophic chromosome damage, losses and rearrangements. Such entanglements and interlocks do occur in diploids but are much more commonly resolved (Storlazzi et al., 2010). Thus, meiotic stability is a key hurdle that must be overcome following WGD and is one of the hallmarks of an adapted polyploid. Indeed, loci that encode genes controlling meiotic recombination and crossovers are strongly implicated in adaptation to WGD (Hollister et al., 2012; Yant et al., 2013).

Of course, one way to bypass unstable meiosis is to simply not use it. It has long been recognized that asexual reproduction (vegetative propagation and agamospermy) and WGD are correlated, with polyploids displaying elevated rates of asexual propagation compared to diploid relatives (Manning and Dickson, 1986; Schinkel et al., 2016; Herben et al., 2017). Such a reproductive strategy may even confer short-term benefits. Asexual reproduction is considered advantageous during range expansion (Baker, 1967), and this may go some way in explaining the invasive nature of many polyploids. Likewise, vegetative propagation could be a short-term fix; buying time for stable meiosis to evolve (Vallejo-Marín and Hiscock, 2016) and reducing the frequency of mating with the diploid progenitors (see section Cytotype and Competitive Exclusion below). Despite potential short-term gains, however, asexuality may not be a viable long-term strategy, and many polyploids are sexual.

In established sexual autopolyploids meiotic instability is often resolved (Yant et al., 2013; Bomblies et al., 2015), with chromosomes forming bivalents or multivalents that segregate regularly. Indeed, there are multiple conceivable ways to modify meiosis and escape genomic instability, but there is considerable empirical work yet to do to learn how many of these exist in nature, much less the mechanistic basis of different solutions. An elegant theory suggests that simply increasing the degree of CO interference could solve this problem (Bomblies et al., 2016). Under this theory if the range of CO interference is greater than the distance to the end of the chromosome, the number of COs would be reduced to one, and if the range of CO interference is only slightly smaller than the chromosome length the COs will be terminalized. This would favor conformations that produce appropriate tension leading to orderly anaphase (Bomblies et al., 2016). Whatever the mechanism, stabilizing meiosis would seem the best solution given the advantages of sex in the long run.

Gene Dosage

It was historically expected that an increase in gene number would result in a uniform increase in gene expression (Comai, 2005). This would correspond to a 1:1 dosage effect where 1*x* diploid expression per genome results in double the total gene expression per cell in an autotetraploid. However, other dosage responses are also possible (Coate and Doyle, 2010): for example, dosage compensation could reduce the per genome expression by half to match overall diploid expression levels per cell (0.5:1 in autotetraploids). This compensation could be partial (ratio between 0.5:1 and 1:1) or even negative (ratios below 0.5:1). In the other direction, dosage effects could also amplify the expression level increase resulting from polyploidy (ratios above 1:1). The impact of these effects could vary across the transcriptome with some gene categories more likely to follow a 1:1 response while others respond differently. The evidence for an overall 1:1 dosage response to WGD in autopolyploids stems primarily from one study of a synthetic polyploid series in maize (Guo et al., 1996). Among the 18 genes followed, most exhibited a 1:1 dosage response, but there were several exceptions and the dosage compensation response of some genes varied over different ploidy levels. Compared to the extensive literature on gene expression changes in allopolyploids (e.g., expression level dominance, genome-wide transcriptomic rewiring, biased fractionation) this represents a major gap in our understanding in autopolyploids. This gap should be closed, as recent empirical evidence points clearly to selective sweeps in transcription-related loci. This suggests that adaptation of the transcriptional machinery to cope with gene dosage effects may be important in neo-autopolyploids. Indeed, one of the most dramatic genome-wide selective sweeps in response to adaptation to WGD in *A. arenosa* is in the locus encoding the *Transcription initiation factor IIF (TFIIF) beta subunit*, a key member of the complex that drives RNA synthesis during the transcription (Hollister et al., 2012; Yant et al., 2013).

This gap in our understanding of the mechanistic basis behind dosage compensation is partly the result of technical difficulties. Methods commonly used to evaluate genome-wide expression patterns (microarrays and RNAseq) rely on extensive normalization of the RNA input and therefore are perfectly appropriate to detect relative changes in gene expression but not at all to measure variations in absolute transcriptome size (Lovén et al., 2012; Coate and Doyle, 2015). Only recently has a directed investigation of expression differences between diploids and autotetraploids been reported, using three normalization procedures to take into account transcriptome size, biomass, and cell density (Visger et al., 2017). This allows a clearer discrimination of expression differences than was previously possible, as concentration-based normalizations can mask up to 50% of expression differences. Indeed, in previous relative transcriptome comparisons (Stupar et al., 2007; del Pozo and Ramirez-Parra, 2014; Zhang et al., 2014) <10% of the transcriptome undergoes expression changes in response to WGD. However, when absolute differences were measured by Visger et al. approximately 1.5 times more genes showed expression differences; 80% of which had a dosage sensitive response with a ratio >0.5:1 (overexpressed in autotetraploid cells compared to diploids) thus compensating for the lower cell density of tetraploids (similar expression per biomass). As the first global analysis of gene dosage response to WGD properly taking into account potential changes in overall transcriptome sizes, this study by Visger et al. effectively demonstrates that 1) most genes are compensating for gene dosage (83% with no differences between diploid and tetraploid cells) and 2) that the genes which do not (17%) mostly present increased expression per cell (ratio >0.5:1) somehow compensating for gene dosage per biomass. This data thus does not support a general trend for 1:1 dosage effects, and also shows other responses are possible (underexpression per cell in tetraploids). More empirical work in this area is required (**Figure 2**) as antagonistic dosage effects in particular may open possibilities for WGD-induced transcriptomic innovations. In particular, some functional categories were enriched for dosage-sensitivity, mostly in relation to photosynthesis or the chloroplast.

Indeed, parallel evidence from the study of patterns of gene retention following allo-polyploidization (Maere et al., 2005; Thomas et al., 2006; Coate et al., 2011) or from more manageable experimental systems like the X chromosome (Birchler, 2014) support the idea that dosage responses are selectively constrained by genetic pathways. This idea was formalized as the dosage balance hypothesis by Papp et al. (2003), who argued that greater fitness loss would result from perturbing the relative abundance of components of a signaling cascade or of a multi-subunit protein complex than from absolute but concerted concentration changes that would maintain overall stoichiometry. For genes under this selection for gene dosage balance, WGD itself may not greatly disrupt relative abundances, as all genes would see their dosage increased more or less proportionally to one another. If that is the case, however, gene dosage compensation mechanisms (from reduction of gene expression to loss of gene duplicates) would be strongly constrained following WGD, as these would need to be concerted across all interacting subunits to maintain stoichiometry. Hence, the gene duplicates of interacting proteins have been well-preserved even over long time-spans following polyploidization (Papp et al., 2003; Thomas et al., 2006; Birchler and Veitia, 2010). This could be a major hindrance for polyploids as this selection would limit an ability to rectify deleterious gene dosage effects. One way to circumvent this selection would be a uniform reduction of gene expressions across the genome, which, Freeling et al. (2015) argue, is precisely one of the effects predicted for transposition bursts (see section Transposition Burst and the Generation of a High-Effect Mutation Pool below). Supporting this potential importance of TEs in dosage response, dosage-dependent genes in *A. thaliana* × *A. arenosa* polyploid series containing from 4 to 0 copies of the *A. thaliana* genome were depleted of TEs while dosage-independent genes were enriched in TEs (Shi et al., 2015). Under the assumption that TEs are equally likely to insert near both categories of genes, this data suggests that TE insertions near dosage-independent genes are selected against. This is consistent with the gene dosage balance hypothesis, as genes within a network would be unlikely affected synchronously, except in the case of a transposition burst where the effect of TEs would be evenly scattered across the genome. This hypothesis remains to be explored further (**Figure 2**), as

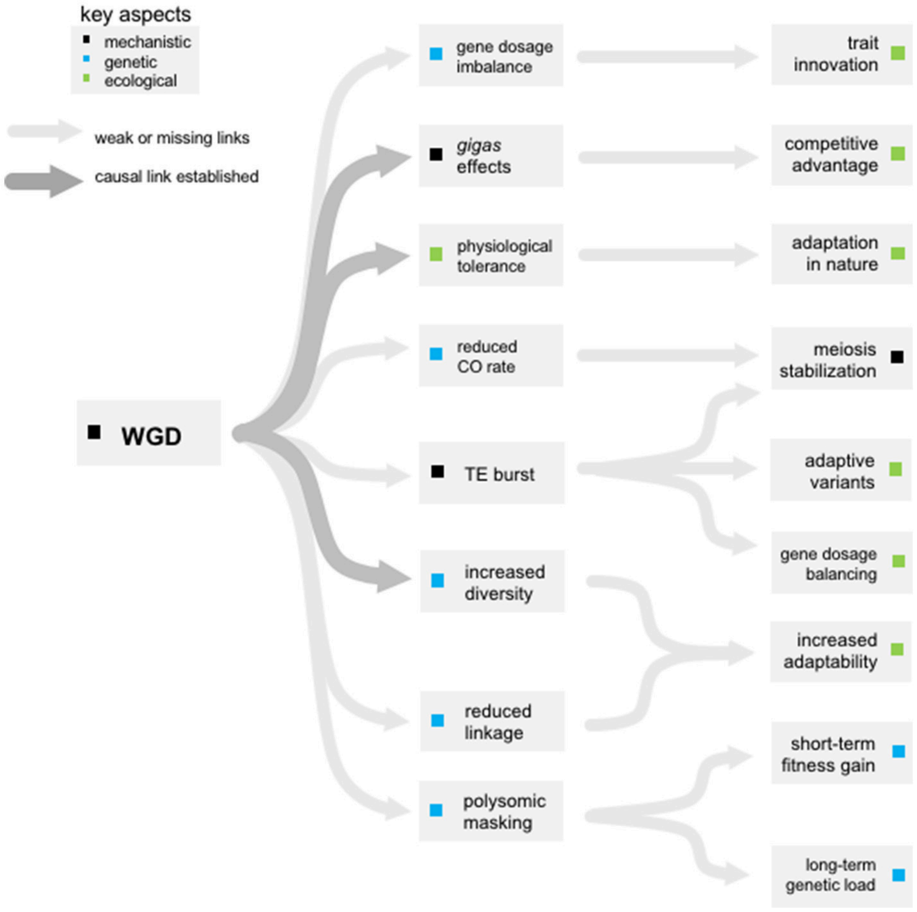


FIGURE 2 | The missing links. The major direct and indirect consequences of WGD either firmly established (dark gray) or still in need of experimental confirmation (light gray lines).

there could also be reverse effects, where presence of TEs in the vicinity of a gene might itself influences the gene-dosage response of that gene. Such effects are yet to be tested but it could help explain the diversity of transcriptomic responses to WGD observed even between accessions of the same species. In *A. thaliana*, Yu et al. (2010) detected ∼500 genes differentially expressed in the Col-0 genotype after WGD but only nine in Ler. Notably, the three genes Yu et al. identify as highly but differentially ploidy-responsive across Col-0, Ler, and a panel of seven other accessions (AT1G53480, AT4G32280, AT5G18030) are all located within 1 kb of a TE insertion (anno-j.org).

Cytotype and Competitive Exclusion

Newly formed polyploids are expected to suffer a mating disadvantage when they are relatively rare in the population (Husband, 2000). This type of frequency-dependent disadvantage is known as minority cytotype exclusion (Levin, 1975). The minority cytotype principle is based on the idea that, under random mating, rare cytotypes are expected to be involved in interploidy matings more often than common cytotypes. Assuming that interploidy matings are more likely to produce inviable or sterile offspring, rare cytotypes should have reduced relative fitness. Such a frequency-dependent mating disadvantage was described from experimental and natural mixed-ploidy populations (Hagberg and Ellström, 1959; Maceira et al., 1993; Husband, 2000; Baack and Stanton, 2005; Mráz et al., 2012), but only a few studies further evaluated its significance for polyploid establishment. Interestingly, studies of mixed-ploidy populations of *Chamerion angustifolium* indicate a surprising asymmetry in this relationship between ploidies. In experimental arrays in field conditions, diploid fitness was frequency-dependent, while fitness in tetraploids was unaffected by their relative frequency. This was likely a result of pollinators preferentially visiting flowers of tetraploid individuals, particularly when rare, and also due to skewed pollen competition favoring tetraploids (Husband, 2000; Husband et al., 2002). Interestingly the effect was mirrored in natural mixed-ploidy populations, where tetraploid mothers produced fewer triploid hybrids than diploid mothers (Husband and Sabara, 2003; Sabara et al., 2013). These studies thus provide a demonstration of minority cytotype exclusion in action and a novel mechanism by which polyploids may avoid its consequences through assortative mating. Indeed, given that WGD can yield larger flowers through the *gigas* effect (Ramsey and Schemske, 2002; Simon-Porcar et al., 2017), and pollinators often show a preference for visiting larger flowers, non-random mating in mixed-ploidy populations may be important for alleviating the costs of rarity. Additional mechanisms to reduce the costs of inter-cytotype mating are to shift toward self-pollination (Barringer, 2007), or bypass sex altogether either through agamospermy (Thompson and Whitton, 2006; Kao, 2007), or increased vegetative reproduction, as shown both by association studies, (Herben et al., 2017), and in synthetic polyploids (Drunen and Husband, 2018).

In addition, direct competition between the parental diploid and its derivative autopolyploid can hinder the establishment of a nascent polyploid as predicted by theory (Rodríguez, 1996; Yamauchi et al., 2004). This of course depends on niche overlap between the diploid and polyploid. Unlike allopolyploids, where hybridization is expected to create novel genetic combinations unique to the hybrids, autopolyploids may not immediately possess such dramatic genetic differentiation from their progenitors. On the other hand, ploidy-altered traits may translate to better polyploid performance in competition either with its diploid progenitor or with other species. Studies experimentally addressing competition between diploids and their naturally-occurring, recently arisen autopolyploid derivatives are, however, very rare and either support this view (Maceira et al., 1993) or show no difference (Thompson et al., 2015). Alternatively, ploidy-altered traits may also help to cope with competition with other species and may broaden niches, opening the possibility to escape from minority cytotype exclusion. This notion is supported by theoretical models (Rodríguez, 1996) and the observation that polyploids are more frequent in competitive, demanding, and human-disrupted habitats than their diploid relatives (Ehrendorfer, 1980). However, despite the frequent invocation of superior competitive ability to explain polyploid success, this has only rarely been addressed experimentally and available results speak against this trend in autopolyploids (Münzbergová, 2007; Fialová and Duchoslav, 2014), in contrast to allopolyploids (Rey et al., 2017).

THE SHORT-TERM GAINS

Masking of Deleterious Mutations

Haldane pointed out in 1933 that in the short-term polyploidy should greatly reduce the effect of genetic load by masking recessive or partially-recessive deleterious mutations behind an increased allelic multiplicity (Haldane, 1933). Indeed, at a given allele frequency q, the proportion of homozygotes in a diploid population will be q^2 but this drops exponentially to q^4 in an autotetraploid population. Thus, for deleterious recessive alleles, the frequency of autotetraploids expressing the associated phenotype will be an order of magnitude lower (or two if q is already small). This means that deleterious recessive alleles can reach much higher frequencies in autotetraploid populations before being exposed to strong selection and equilibrium frequencies are higher in autotetraploids vs. diploids. As newly formed polyploids initially inherit genetic load from a diploid genomic background where the equilibrium frequency is much lower, genetic load will be relieved in young polyploids, providing an early benefit (**Figure 1**). As a result, as long as most deleterious alleles are at least partially recessive (which is the case in both *Mimulus* and yeast; Willis, 1999; Agrawal and Whitlock, 2011), WGD is predicted to lead to temporary fitness increases (Korona, 1999; Otto and Whitton, 2000; Mable and Otto, 2001), although empirical evidence for this is still missing (**Figure 2**).

Instantly Altered Physiological Properties

Both population genetic theory and emerging empirical evidence suggest that a broad set of factors interact to alter the genomic landscape of autopolyploids. However, understanding the effect of WGD in isolation from separate yet correlated effects has only recently made major progress. While it was suggested 35 years ago that biochemical and physiological changes resulting from WGD might underlie polyploid adaptability (Levin, 1983), the best evidence of a direct link took three decades to emerge, when Chao et al., 2013) elegantly demonstrated that *A. thaliana* first generation autotetraploids have instantaneously enhanced salt tolerance compared to isogenic diploids. Neo-autotetraploid *A. thaliana* lines were shown to experience a tradeoff, being less fit compared to diploid progenitors under non-saline conditions, but more fit in response to saline challenge (Chao et al., 2013). The authors proposed that in conditions of salinity stress the autopolyploid lineages would benefit from a fitness advantage that could contribute to their establishment and persistence, thanks to an improved ability to accumulate potassium and exclude sodium. Indeed, the following year it was shown that autotetraploid *A. thaliana* are additionally more drought tolerant (del Pozo and Ramirez-Parra, 2014). A major challenge now is to determine the molecular events that bind WGD to this enhanced stress tolerance. It appears that the key tissue to investigate is the root, where salinity and drought tolerance meet potassium homeostasis and ABA signaling (Saleh et al., 2008; Meng et al., 2011; Allario et al., 2013; Chao et al., 2013; Wang et al., 2013; del Pozo and Ramirez-Parra, 2014). Work there promises to reveal mechanistically how WGD has an immediate effect on cellular physiology that is independent of increased genetic diversity.

There is also good evidence that somatic WGD may enhance stress resilience. For example, in *Medicago* and sorghum root endopolyploidy correlates with salt tolerance (Ceccarelli et al., 2006; Elmaghrabi et al., 2013) and can be induced by salt in tolerant, but not sensitive, strains (Ceccarelli et al., 2006). Thus, the ability to induce endopolyploidy may be responsible for salinity tolerance, perhaps due to cell size changes in the roots that alter ion uptake. Higher proportions of endopolyploid cells also correlate with greater drought tolerance (Cookson et al., 2006; Saleh et al., 2008; Meng et al., 2011; Chao et al., 2013).

Equally importantly, some effects have been disentangled from polyploidy and shown to be unrelated. A recent study by Solhaug et al. (2016) demonstrated that the allopolyploid *Arabidopsis suecica* had enhanced carbon assimilation via photosynthesis and elevated respiration rates relative to its progenitors *A. arenosa* and *A. thaliana*. This enhanced photosynthetic capacity was environment specific (dependent on high light levels) suggesting a potential mechanism for range expansion by the allopolyploid into novel niches. This advantage was not the direct result of polyploidization, as shown by comparing 12 accessions of isogenic diploid *A. thaliana* to colchicine-generated neo-polyploids. These autopolyploids showed no difference in carbon assimilation by photosynthesis compared to their diploid progenitors, suggesting that the photosynthetic vigor of *A. suecica* is a result of hybridization and not WGD.

Despite this progress, the majority of functional studies do not capture the final link: proof that the observed WGD-associated change is adaptive in the natural environment (**Figure 2**). An exception was provided by Ramsey, who used reciprocal transplants involving tetraploids, hexaploids, and neo-hexaploids (produced from the tetraploids) of *Achillea borealis* to show a link between WGD and increased fitness in the native environment (Ramsey, 2011). There, WGD itself accounted for 70% of the fitness difference, while the remaining variation (i.e., difference between neo-hexaploids and native hexaploids) could be ascribed to subsequent evolution of the native polyploid. However, the physiological mechanism and its genetic basis in this case remains unknown, which highlights the difficulty of comprehensive inter-disciplinary studies combining genetics, physiology, and ecology.

Transposition Burst and the Generation of a High-Effect Mutation Pool

The hypothesis that WGD presents a genomic shock that activates transposable elements (TEs) across the genome was first proposed by Barbara McClintock (1984) to explain the association between polyploidy and increased TE content. Resident in virtually all genomes, TEs are highly mobile, making them powerful endogenous mutagens. To repress their activity, organisms target TEs with epigenetic silencing mechanisms such as DNA methylation (Bennetzen and Wang, 2014; Ito and Kakutani, 2014). However, the efficiency of TE silencing can be influenced by a number of factors, including environmental or cellular stressors. In some cases, the reactivation of TEs can be explained by the presence of stress-associated transcription factor binding sites in TE promoters (reviewed by Horváth et al., 2017). However, a more global impact of stress on the efficiency of TE-silencing mechanisms has also been suggested (Tittel-Elmer et al., 2010). In particular, genomic stress brought on by hybridization or polyploidization has global effects on epigenetic regulation and may thereby lead to TE reactivation (Kashkush et al., 2003; Madlung et al., 2005; Lopes et al., 2013; Springer et al., 2016; Edger et al., 2017). However, genome shock in polyploids has been studied primarily in allopolyploid contexts, where hybridization is the major contributor, as observed in *Senecio cambrensis* (Hegarty et al., 2006) and *Spartina anglica* (Parisod et al., 2009). To date, very few studies addressed the effect of WGD *per se* on TE transpositions apart from Bardil et al. (2015) who demonstrated an activation of LTR-retroelements following WGD, along with a contribution of gene-flow at the origin of polyploids.

Although most of the mutations TEs generate are deleterious, there is some evidence that TE insertions can be beneficial. The best example of this adaptive potential can be found in the classic case of industrial melanism in the peppered moth, where a young TE insertion that appeared and rapidly rose to fixation during the industrial revolution (∼200 years ago) has been proposed as responsible for the dark morph providing camouflage from predators (Van't Hof et al., 2016). The variation produced by TE activity can thus become a fruitful target of natural selection, providing adaptive solutions to the very stresses that initiated their reactivation (Ito et al., 2016). Thus, a global transposition burst triggered by genomic shock could immediately provide nascent polyploids with a pool of high-effect mutations to test against new challenges. In addition, the reactivation of TEs in young polyploids may also contribute to the stabilization of the neo-polyploid genome. First, TE insertions close to genes are known to have an impact, mostly negative (Hollister and Gaut, 2009) but sometimes positive (Quadrana et al., 2016), on the expression of nearby genes. Therefore, the global array of new insertions resulting from a transposition burst might result in broad re-wiring of gene expression and thereby contribute to the rebalancing of gene-dosages (Kashkush et al., 2003; Freeling et al., 2015), as was suggested by recent observations in rice neopolyploids (Zhang et al., 2015; see section Gene Dosage above). Second, TE content in centromeric regions contributes to the bulk of centric heterochromatin that is essential for the separation of sister chromatids during meiosis. Heterochromatin resists the pull exerted by microtubules and the resultant tension silences the spindle checkpoint, allowing meiosis to proceed (Stephens et al., 2013). Increased TE content generated from a transposition burst in neo-polyploids and distributed across chromosomes may thus lead to an overall strengthening of the meiotic spindles and contribute to stabilizing chaotic meiosis following WGD (see Meiosis section above).

THE LONG-TERM GAINS

Enhanced Invasiveness and Colonization Potential

Polyploids are over-represented among invasive plants. While in many cases diploids and tetraploids co-exist in the native range, the tetraploids are more often found alone in the invaded range than the contrary (e.g., Hollingsworth and Bailey, 2000). Consistent with this, polyploidy is associated with a potential for habitat colonization and transitions to weediness (Brown and Marshall, 1981; Soltis and Soltis, 2000; Pandit et al., 2006, 2011; Prentis et al., 2008). Common physiological factors contributing to invasiveness are associated with the necessary tolerance to environmental variation including stress resilience, phenotypic plasticity, or rapid cycling (early and prolific flowering aids in coping with or escaping from unpredictable environmental conditions; Baker, 1965; Grotkopp et al., 2002; Blair and Wolfe, 2004; Burns, 2004; Hall and Willis, 2006; Sherrard and Maherali, 2006; Franks et al., 2007). Such life history adaptations can help mediate trade-offs between resource accumulation and stress-avoidance and are important for wild species as well as for crops (Jung and Müller, 2009).

Are polyploids pre-adapted or innately more capable of acquiring such traits? This is an open question, but the many cases where both cytotypes occur in their native range but only polyploids do in the invasive ranges (Lafuma et al., 2003; Mandák et al., 2005; Kubátová et al., 2007; Schlaepfer et al., 2008; Treier et al., 2009) suggest a potential pre-adaptation of polyploids for invasiveness (te Beest et al., 2012). However, environmental stresses also increase the rate of unreduced gamete formation and thus of polyploidization events (Bretagnolle and Thompson, 1995; Ramsey and Schemske, 1998). Therefore, polyploidization has also been viewed as a post-colonization process (Mandák et al., 2003) even if through hybridization (e.g., Hahn et al., 2012). Here we focus on the genetic factors implicated in invasiveness that are likely impacted by WGD (**Figure 1**). In particular, because the invasion of novel habitats typically proceeds from a small number of founders, some genetic properties of autopolyploids can enhance their chances of successful colonization. These include larger effective population sizes, a greater tolerance for selfing (and inbreeding depression), the ability to recover from the genetic bottlenecks, potentially enhanced sampling from existing standing variation, as well as expected lower levels of linked selection (below and te Beest et al., 2012).

Increased Diversity and Tolerance for Selfing

In allopolyploids, where two distinct genomes are united, fitness advantages have often been attributed to interspecific hybridization rather than WGD (Barker et al., 2016). A conservative back-of-the-envelope calculation by Barker et al. (2016) estimated that the rate of production of autopolyploid cytotypes could be 40–80 times greater than that of allopolyploids. Given the approximate parity of allo- and auto-polyploids in nature, this suggests a large advantage to hybridization over the benefits directly attributable to WGD.

Because the two sub-genomes typically do not recombine, allopolyploids can continue to enjoy the advantage of heterosis and a stable multi-allelic state over many generations. Autopolyploids on the contrary, do not benefit from fixed heterozygosity. Nevertheless, it has been proposed that polysomic inheritance alone leads to higher genetic diversity (Haldane, 1932), and experimental comparisons between autotetraploids with tetrasomic inheritance and their diploid parents validate this theoretical expectation (Soltis and Soltis, 2000). This increased diversity has been linked to both an immediate increase in the number of mutational targets (doubled number of chromosomes in the case of autotetraploids), that in the long-run provide increased effective population sizes, and an expected reduced efficiency of purifying selection (Ronfort, 1999). This rise in genetic diversity in autopolyploids is proposed as a cause of the observed successes of tetraploids compared to their diploid sister lineages (Roose and Gottlieb, 1976; Soltis and Soltis, 1989, 2000; Soltis et al., 1993; Brochmann et al., 2004). The positive relationship between phenotypic plasticity and invasiveness introduced first by Baker (1965) is now well-documented through numerous physiological and morphological comparisons of invasive and native species (reviewed in Richards et al., 2006). Increased diversity in polyploids is often invoked to explain an increased plasticity and ability of polyploids to sustain range expansions into disturbed habitats. This is a tempting speculation, but a causal link demonstrating that the increased diversity in tetraploids confers an adaptive advantage is lacking (**Figure 2**).

In addition, reduced homozygosity in autopolyploids is expected to reduce the potential inbreeding depression associated with genetic load (Charlesworth and Charlesworth, 1987). This is because at any locus the increase in copy number in autopolyploids increases the probability of heterozygous offspring, even during selfing (Moody et al., 1993). As a result, the fitness cost associated with selfing (inbreeding depression) may be ameliorated (Lande and Schemske, 1985; Schemske and Lande, 1985) or at worst unchanged (Ronfort, 1999) depending on the range of dominance effects impacting fitness. This prediction has been confirmed in ferns, where the self-fertilization of the gametophyte makes it possible to directly measure the impact of selfing on survival rates in the resulting sporophytes. In two different diploid-tetraploid fern pairs, selfing survival rates were nearly 100% in the tetraploid races, while it ranged from 5 to 60% in the diploids (Masuyama and Watano, 1990). Similarly, a reduction of inbreeding depression is observed in other polyploid-diploid comparisons (Husband and Schemske, 1997; Galloway et al., 2003; Husband et al., 2008), even though there are cases where the opposite is observed (Johnston and Schoen, 1996).

Tolerance to selfing is of major importance in the ability of a population to colonize new habitats, a consideration known as "Baker's rule" (Baker, 1967). Indeed, during the colonization process, early invaders are likely to be isolated with little opportunity for outcrossing. Selfing therefore provides reproductive assurance for the dispersed invaders (Barrett et al., 2008), and this translates to a high rate of co-occurrence between selfing or asexual propagation and low-density conditions or frequent colonization bouts (Baker, 1967; Price and Jain, 1981; Pannell and Barrett, 1998). For example, Daehler found that low inbreeding depression in hexaploid smooth cordgrass populations invading the San Francisco Bay area in California was associated with higher self-fertility and a higher fitness advantage for founding populations in the field (Daehler, 1998; Renny-Byfield et al., 2010).

Reductions in Linked Selection: An Advantage in Changing Environmental Conditions?

As a mirror image to the reduced efficiency of selection against deleterious mutations, the increase in frequencies of beneficial alleles, even when dominant, will be slower in polyploids under tetrasomic inheritance than in diploids (Hill, 1971). Therefore, the time to fixation for an allele during a selective sweep can be greatly increased in autopolyploids. This prolonged rise in allele frequency might lead to more opportunities for mutation and recombination with other haplotypes, which are even further enhanced by the increased mutation and recombination rates resulting from greater ploidy levels. Weaker linkage thus may promote adaptation through reduced interference between alleles, allowing greater opportunity for a beneficial allele to recombine onto haplotypes with fewer deleterious mutations (**Figure 1**). However, increased recombination can lead to lower fitness in constant environments by breaking down beneficial associations (Lewontin, 1971; Feldman et al., 1980). Therefore, increased recombination may only be selected for in environments with fluctuating conditions (Charlesworth, 1976; Otto and Michalakis, 1998; Lenormand and Otto, 2000), which also happen to be environments with higher incidences of polyploids (Favarger, 1984; Brochmann et al., 2004; Parisod et al., 2010). This association between increased recombination and adaptation to environmental variation would strongly favor long-term evolution of autopolyploids, but remains to be experimentally tested (**Figure 2**).

Sampling of Standing Variation From Local Introgression

As an autopolyploid lineage expands its range, it may encounter populations of its diploid progenitor or other species with which hybridization is possible. Provided that such populations are locally adapted, introgression may then supply genetic variants that facilitate persistence. Although polyploidization is traditionally viewed as a means of instant speciation (Coyne and Orr, 2004), the ploidy barrier is permeable (reviewed in Ramsey and Schemske, 1998; Kolár et al., 2017). While adaptive introgression is increasingly recognized as an important force in the evolution of haploid and diploid organisms by genomic studies (reviewed by Arnold and Kunte, 2017; Schmickl et al., 2017), empirical genomic evidence for gene flow among a diploid and its autopolyploid derivative is lacking. The ability to accept genetic variation from alternative cytotype might be beneficial, as it could provide preadapted local alleles upon which selection may act and/or may alleviate inbreeding associated with founding events during range expansions (Parisod et al., 2010). We however lack well-documented examples of traits and underlying loci that may explain evolutionary significance of gene flow for establishment and further spread of a polyploid. The only case to our knowledge, although confined to allopolyploids, is across-ploidy transfer of potentially adaptive floral genes, RAY1 and RAY2, from diploid *Senecio squalidus* into the allotetraploid *Senecio vulgaris* that has given rise to a novel variety of *S. vulgaris* with ray florets (Chapman and Abbott, 2010). An additional hint, coming from an autopolyploid system, is the likely uptake of a diploid-like *CONSTANS* allele during the colonization of railways by a distinct lineage of autotetraploid *A. arenosa* (Baduel et al., 2018). This allele may allow the railway ecotype to escape the repression exacted on flowering by *FLOWERING LOCUS C* and underlie the observed rapid and repeated flowering. These two examples indicate that this may be a fruitful area for future research. An alternative benefit of interploidy hybrids for polyploid establishment may result from their contribution to recurrent formation of polyploids. Triploid hybrids, if fertile, often produce unreduced ($2n = 2x$) gametes (Ramsey and Schemske, 1998; Chrtek et al., 2017) that can merge with reduced ($n = 2x$) gametes of a tetraploid leading to formation of novel tetraploids (Husband, 2004). We note, however, that much gene flow may be neutral or even maladaptive. For example, if a tetraploid has adapted to problems associated with meiotic segregation during its establishment (Yant et al., 2013), later diluting of such co-adapted gene networks by introgression of diploid-like alleles would lead to reductions in fitness.

Even when assuming beneficial consequences, interploidy gene flow would provide relative advantages to the polyploid only in cases when (potentially adaptive) alleles flow more often into the polyploid than into progenitor diploids (**Figure 1**). Indeed, this seems to be the case and it was recognized as early as 1971 by Stebbins that gene flow among cytotypes is asymmetrical (Stebbins, 1971). A mechanistic explanation for this asymmetry is that while there are direct pathways for gene flow in the $2x$ –> $4x$ direction, the reverse is more convoluted. The unreduced $2n = 2x$ gamete of a diploid and the reduced $n = 2x$ gamete of a tetraploid can combine leading to one-step formation of a tetraploid interploidy hybrid ($2x + 2x = 4x$; Koutecký et al., 2011; Chrtek et al., 2017; Sutherland and Galloway, 2017). However, a triploid hybrid, capable of forming reduced $n = x$ gametes, is an essential stepping-stone for the creation of a diploid hybrid (Kolár et al., 2017). Thus, gene flow in the $4x$ –> $2x$ direction is less likely as it involves two separate crossing steps ($4x$ –> $3x$ and $3x$ –> $2x$). Moreover, the triploid hybrid is often either non-viable (triploid block) or unfit (Ramsey and Schemske, 1998; Köhler et al., 2010). Indeed, the few available empirical genetic studies document either much stronger (in autopolyploid systems: Ståhlberg, 2009; Jørgensen et al., 2011; Arnold et al., 2015) or exclusively unidirectional gene flow from the diploid into the polyploid (in allopolyploid systems: Slotte et al., 2008; Chapman and Abbott, 2010; Zohren et al., 2016). In a longer evolutionary timespan recurrent origins of autopolyploid lineages from different diploid sources followed by hybridization among these polyploids (Soltis and Soltis, 2009) would also lead to enrichment of the tetraploid gene pool by alleles from distinct diploid lineages, similar to direct unidirectional gene flow from diploid to polyploid.

If higher polyploids are formed (i.e., hexaploids, octoploid, etc.) they may hybridize with the tetraploid or among one another and further enhance variation of the polyploid lineages. The few empirical studies available show that the postzygotic barrier, both in terms of rate of hybrid formation and its fitness, is lower among the various polyploid cytotypes than it is between diploids and their polyploid derivatives (Greiner and Oberprieler, 2012; Sonnleitner et al., 2013; Hülber et al., 2015; Sutherland and Galloway, 2017). This corresponds well with the explanation of maternal: paternal genome imbalance in the endosperm as a primary cause of the postzygotic barrier (Köhler et al., 2010; Greiner and Oberprieler, 2012). Because the magnitude of endosperm imbalance in tetraploid–hexaploid hybrids is approximately one third lower than in diploid–tetraploid hybrids (Sonnleitner et al., 2013) these higher-ploidy hybrids may be more fit than diploid–tetraploid hybrids.

Aside from intraspecific gene flow, polyploidy may also break down systems of reproductive isolation present in diploid progenitors and thus increase interspecific gene flow. For example, although the reproductive isolation in diploid lineages of *Arabidopsis arenosa* and *Arabidopsis lyrata* is near complete, tetraploid *A. lyrata* can form viable hybrids with both diploid and tetraploid *A. arenosa,* likely due to the disruption of an endosperm-based barrier (Lafon-Placette et al., 2017).

Interestingly, hybridization between those species appears to have donated beneficial alleles contributing to local adaptation to harsh serpentine soils in the tetraploid *A. arenosa* (Arnold et al., 2016). In this study Arnold et al. (2016) found that several genes exhibiting signatures of selection for adaptation to serpentine soils also appeared to have been introgressed from *A. lyrata*. Finally, the tendency of polyploids to expand into novel niches may further increase chances of encountering foreign lineages with which hybridization may occur. Although the cause of this is unclear, the heightened adaptability of many polyploids fueled by introgression may provide positive feedback, allowing further spread and hybridization. Altogether, these examples illustrate a tendency for polyploids to act as evolutionary "sponges," accumulating variation through introgression across both ploidy and species barriers. This supports the view of polyploids as diverse evolutionary amalgamates from multiple distinct ancestral lineages—a property advantageous for further expansions.

THE LONG-TERM CHALLENGES

This begs the question: if WGD events are common, and polyploids display advantageous traits, why are established autopolyploids relatively uncommon and paleo-polyploids so frequent? Transitions to polyploidy tend to be observed at the tips of phylogenies (Escudero et al., 2014), suggesting that polyploid lineages typically do not survive as such over longer evolutionary timescales. Consequently, the growing consensus is that polyploidy is an ephemeral but repeatedly appearing state (Wendel, 2015). This could be the result of both pervasive polyploid extinction, as there is some suggestion that recently arisen polyploids experience lower diversification rates and higher extinction rates relative to congeneric diploids (Mayrose et al., 2011, 2015; Arrigo and Barker, 2012), as well as repeated returns to diploidy and disomic inheritance after transitionary polyploid phases (e.g., Haufler, 1987; Wendel, 2015; Soltis et al., 2016). To date, such a transition has only been mathematically modeled in autopolyploids (Le Comber et al., 2010); empirical evidence is lacking. Thus, in addition to the short term biological challenges faced by newly-arisen polyploids, longer-term challenges may help explain the transience of the polyploid state, even reviving the idea of polyploidy as an evolutionary "dead-end" (Wagner, 1970; Stebbins, 1971; Mayrose et al., 2011, 2015). Ironically, many of these postulated longer-term negative effects result from the continuation of earlier beneficial population genetic mechanisms.

Increased Genetic Load

If in the short-term polysomic masking results in a fitness increase, a reduced strength of purifying selection (Ronfort, 1999) would eventually lead to the slow rise of recessive deleterious mutations until mutational load reaches a new, higher equilibrium (Otto and Whitton, 2000). This new polyploid equilibrium may take hundreds of thousands of generations to establish (Otto and Whitton, 2000), but would ultimately produce a genetic load proportional to ploidy level and the mutation rate μ per haploid genome (Haldane, 1937). A particularly strong effect of this would be on TEs, as their distribution of fitness effects is much more heavily skewed toward highly-deleterious mutations compared to single-nucleotide polymorphisms and thus are strongly affected by purifying selection. This has been demonstrated recently in *A. thaliana* (Quadrana et al., 2016), where it was shown that TEs insert throughout the genome, but are rapidly purged from genic rich regions and chromosome arms, most likely due to the deleterious consequences of insertions near or within genes (Quadrana et al., 2016). Even if evidence of this mostly comes from allopolyploid systems (wheat, *Brassica, etc.*), these long-term effects are likely to be similar in autopolyploids and in the long run we can expect the initial differences in transposition burst triggered by the two modes of polyploidization to be less important compared to the relaxation of purifying selection shared by both systems. Indeed, in the allotetraploid *Capsella bursa-pastoris,* an increase in TE content was observed around genes compared to its two parental diploid species, *C. grandiflora* and *C. orientalis*, which was attributed primarily to a relaxation of purifying selection and not to any change in TE activity (Ågren et al., 2016), and there is accumulating evidence of TE proliferation over long timespans following polyploidization (Sarilar et al., 2011; Yaakov and Kashkush, 2012; Piednoël et al., 2015). However, this doesn't seem to apply to all TE families equally. For example some gypsy-like retro-elements proliferated in *Aegilops* tetraploids while others remained quiescent (Senerchia et al., 2014). This could be due to differences in insertion preferences between TE families or more simply to the fact that many TEs are actually defective. Indeed, most of the TEs carried by a genome have lost their transpositional capacities and are fossilized: in the human genome <0.05% of TEs remain active (Mills et al., 2007). Between two active families, differences in their regulation, copy number, chromosome localization, etc. may also explain different responses to relaxed purifying selection. For example, a family inserting more commonly into genes will be more strongly purified and therefore more strongly affected by a relaxation of purifying selection than in TE families that inherently avoid inserting into gene-coding loci. Such differences in insertion preferences have been observed in one LTR retrotransposon family between *A. thaliana* where genic insertions are strongly selected against, and *A. lyrata,* where gene-poor centromeric regions are preferentially targeted, reaching much higher copy numbers (Tsukahara et al., 2012).

Eventually, it is thought that the reduced strength of purifying selection from polysomic masking may overshadow the early advantages of low mutational load, which begins at the lower diploid equilibrium levels immediately following WGD (Otto, 2007; Gerstein and Otto, 2009). At equilibrium (**Figure 1**), polyploids are predicted to suffer from the increased frequency of deleterious mutations, which are introduced in higher numbers (doubled in the case of autotetraploids). However, given the difficulties of finding an ancient enough system where autopolyploids have reached such an equilibrium but are still ecologically comparable to their diploid progenitors, empirical support for this remains sparse (**Figure 2**).

Slower Selection on Recessive Beneficial Mutations

In addition to hampering selection against deleterious mutations, polysomic masking can also prevent recessive beneficial mutations from reaching fixation. This may even effectively counter the increased input of beneficial mutations arising from the increased number of haploid genomes (Haldane, 1932; Gerstein and Otto, 2009). Whereas in haploids the rate of fitness increase only depends on the rate of appearance of beneficial mutations and their fitness effect (Haldane et al., 1927), in diploids it also depends on the dominance level of mutations. This is further intensified in polyploids (Gerstein and Otto, 2009). For example, in autotetraploids with tetrasomic inheritance, the rate of fitness increase (w) can be written as a function of the rate of appearance, the fixation probability, and the fitness effect (s) as follows:

$$\Delta w_{4x} = 4N\nu.2h_1 s.s$$

Where N is the population size, ν is the beneficial mutation rate, and h_1 is the dominance of the new allele in simplex (for example Aaaa for tetraploids). Therefore, polyploids would adapt faster only when mutations are at least partially dominant ($h_1 > 0.5$) and thus not hindered by polysomic masking. In an attempt to test this prediction experimentally, Schoustra et al. (2007) observed the fastest rates of loss of a costly resistance allele in diploid strains of the fungus *Aspergillus nidulans* that periodically reverted to haploidy. These strains accumulated multiple recessive beneficial mutations in the diploid state that were exposed to positive selection in the haploid state. This pattern is reminiscent of the transitionary polyploid phases postulated to have occurred throughout the evolution of plants (Haufler, 1987; Wendel, 2015; Soltis et al., 2016).

Further amplifying this effect of reduced positive selection, lower linkage in autopolyploids (see Reductions in Linked Selection: An Advantage in Changing Environmental Conditions?) increases the likelihood of recombination breaking down favorable haplotypes as they slowly rise in frequency. As a consequence, beneficial mutations in close vicinity and positively selected in autotetraploids are unlikely to remain linked to each other for long. This may be beneficial in the early stages of invasion (directional selection) or under fluctuating environments, but in the long run it is predicted to be unfavorable. Indeed, once established in their new range and closer to a new fitness optimum, selection is predicted to favor increased linkage and reduced recombination (Feldman et al., 1980, 1996).

Bigger Genomes, Slower Growth Rates

With their doubled genomes, autopolyploids are likely to face the general rule in animals and plants dictating that increased genomic content results in decreased growth and division rates (Gregory and Mable, 2005; Otto, 2007). If the impact of genome size is clear at the cellular level it is less evident at the organism level (Knight and Beaulieu, 2008) and exceptions to this rule can easily be found: first in the growth form (e.g., trees have small genomes, Beaulieu et al., 2008) and in the environment (Zörgö et al., 2013). This led some to suggest the overall negative relationship between genome size and metabolic rate across gymnosperms and angiosperms may be the result of a rather indirect effect through other traits such as growth form (Beaulieu et al., 2007a). It should be noted, however, that these rules are based on the observation of established polyploids and therefore the impact of genome size itself remains to be directly assessed independently of the potentially confounding effects of subsequent evolution. On the cellular level, it seems clear that increased nuclear content leads to increased cell volume (Beaulieu et al., 2008; Knight and Beaulieu, 2008) and slower growth rates (Cavalier-Smith, 1978; Gregory, 2001), which have long been observed in polyploids as well (Müntzing, 1936; Stebbins, 1971; Garbutt and Bazzaz, 1983). At the organismal level, older observations have illustrated that polyploids often flower later (Smith, 1946) and are more frequently perennial (Hagerup, 1932; Müntzing, 1936; Sano et al., 1980), but the role of WGD itself has been rarely experimentally evaluated since then. For example, in synthetic *A. thaliana* tetraploids, there was no consistent trend in flowering time over 12 ecotypes investigated in two common gardens (Solhaug et al., 2016) and similarly no differences in this trait were found between diploid and synthetic polyploids of *Chamerion angustifolium* (Husband et al., 2016). On the other hand, a recent study leveraged parallel altitudinal clines and intraspecific genome size variation in maize landraces to show repeated reductions in genome size in high-altitude populations most likely via selection on flowering time (Bilinski et al., 2018). Furthermore, in growth chamber experiments Bilinski et al. were able to confirm an association, even if modest, between genome size, cell production, and cell sizes. Therefore, such constraints may turn out to be particularly costly for polyploids that successfully switched to invasiveness thanks to early advantages (see section Enhanced Invasiveness and Colonization Potential above). Indeed, invasive species commonly exhibit early flowering (Pyšek et al., 2009), lower seed sizes with higher dispersal abilities and annual life cycles which are also the prerogative of small-genome species (Grotkopp et al., 2002; Knight et al., 2005; Beaulieu et al., 2007b). Even if more research is needed to clarify the direct impact of polyploidy, evidence so far suggests that the potential slowing of growth rates may impact negatively long-term fitness. Thus, selection would likely push for a reduction of genome size, especially in transitions to invasiveness. This process may be very long and stochastic, however, as evidence in the *Nicotiana* genus shows genome downsizing is minimal in young polyploids (∼200,000 years old) only appearing in polyploids approximately 4.5 million years old, at which point genome size increases are also observed (Leitch et al., 2008).

Post-polyploidy Diploidization, a Cradle for Diversification

It is now clear that nearly all plant lineages are paleo-polyploids, with their evolutionary histories including at least one round of WGD (Wendel, 2015). However, numbers of past WGD events do not correlate with chromosome numbers nor genome sizes.

For example, given the three rounds of genomic multiplications that have occurred in *Brassica* genomes (α, β, and γ, Franzke et al., 2011; Jiao et al., 2011), and assuming ancestral angiosperms had between 5 and 7 chromosomes (Stebbins, 1971; Raven, 1975), we would expect, without reductions in chromosome numbers along the way, resultant species to carry between 40 and 56 chromosomes, when some carry as few as six (Anderson and Warwick, 1999). A similar reasoning holds with genome sizes (Wendel, 2015): thus it is apparent that past polyploidization events were followed by massive genome downsizing, both in chromosome numbers and in absolute size (Leitch and Bennett, 2004; Leitch and Leitch, 2008). This genome downsizing ultimately leads to the diploidization of descendants (Soltis et al., 2015). These paleo-polyploids then commonly undergo further rounds of polyploidization, generating a cyclical process described as the "wondrous cycles of polyploidy" and occurring repeatedly over long evolutionary timescales (Wendel, 2015).

Several mechanisms have been proposed to underlie the diploidization process, all of them relying on non-homologous translocations (Mandáková and Lysak, 2018). One contributor to these illegitimate recombination events are TEs, since homology between TE copies can lead to spurious recombination events between non-homologous chromosomes (Vicient and Casacuberta, 2017). Some of the rearrangements resulting from these non-homologous recombinations (inversions, reciprocal translocations, deletions, and duplications) do not affect chromosome numbers, but others (end-to-end translocations, EETs, nested chromosome insertions, NCIs, and Robertsonian translocations) are seen as the mechanistic basis of "polyploid drop" (Mandáková and Lysak, 2018). Indeed, all three processes result in the merger of two chromosomes into one via non-homologous recombinations between two distal regions (EETs), two distal regions with a pericentromeric region (NCIs), or between a distal and a pericentromeric region (Robertsonian translocations). Distal and pericentromeric regions are particularly prone to ectopic homologies due to their enrichment in repetitive elements, in particular TEs (Quadrana et al., 2016; Vicient and Casacuberta, 2017). Therefore, even though most recombination events between TEs will lead to small indels, the possibility of large-scale chromosomal rearrangements may represent a major driver of genome restructuring during diploidization (Vicient et al., 1999). In fact, evidence supporting a role for TEs during diploidization has been observed in *Nicotiana* (Lim et al., 2007) and maize (Bruggmann et al., 2006). However, these dysploidy events have an immediate fitness cost, as the merging of two chromosomes leads to obvious chromosome segregation issues. In outcrossers in particular, the probability of forming non-aneuploid offspring is very low, and newly formed dysploids are likely to suffer from woes similar to newly-formed autopolyploids (Mandáková and Lysak, 2018). This is why it was theorized that the establishment of dysploids would be relatively favored in selfers (Charlesworth, 1992). By increasing homozygosity of the offspring, selfing indeed reduces the fitness cost of dysploidy by increasing the probability of producing offspring homozygous for the merged chromosome. Extending this reasoning, we can expect higher rates of dysploidy among weedy invasives, due to both their propensity for selfing and their often faster cycling (e.g., Grant, 1981), which increases the probability of spurious recombinations (Mandáková and Lysak, 2018). This relationship between life-history and dysploidy rate has been confirmed (e.g., Luo et al., 2015) even though some examples show this is not always straightforward (slow polyploid drop rate in rice despite being annual, Murat et al., 2010). Furthermore, the advantage of a reduced number of chromosomes may be particularly valuable for colonizers (see section Bigger Genomes, Slower Growth Rates above).

These considerations become particularly relevant for aging polyploids, which both carry an increased TE content (see sections Transposition Burst and the Generation of a High-Effect Mutation Pool and Increased Genetic Load above) and are more likely to tolerate selfing (section Increased Diversity and Tolerance for Selfing above). Thus, factors that initially represented an advantage for the establishment of recent autopolyploids may transform into the very drivers of polyploid drop and return to diploidy (**Figure 1**).

Compared to WGD, which leads to an exact doubling of chromosome numbers, polyploid drop is more erratic and can produce broad variation in chromosome number. In *Brassica* for example, the variation in base chromosome numbers is the result of multiple and independent diploidizations from the mesohexaploid ancestor (Lysak et al., 2007; Mandáková et al., 2017). Indeed, the stochasticity of polyploid drop, not WGD, is thought to be a major contributor to speciations and radiations (**Figure 1**). However, polyploidy drop is of course not possible without an earlier WGD. Accordingly, recently revised phylogenetic evidence convincingly supports the occurrence of WGDs significantly before large angiosperm radiations, sometimes by millions of years (Tank et al., 2015; Clark and Donoghue, 2017). These reports strengthen the WGD Radiation Lag-Time Model formalized by Schranz et al. (2012), who found that in six examples of angiosperm radiations a species-poor sister-group shared a WGD event with the species-rich crown group, directly contradicting the notion that WGD was the sole immediate cause of these radiations. The lag between WGD and subsequent radiations thus has been proposed as evidence that the long and stochastic process of polyploid drop is the proximal engine of speciation and cladogenesis (Dodsworth et al., 2016; Clark and Donoghue, 2017; Mandáková and Lysak, 2018).

CONCLUSION

As best expressed by Johnathan Wendel, the "wondrous cycles" of polyploidy have gained increasing attention and support, both theoretical and empirical, over the earlier ideas that polyploids were evolutionary dead-ends. Excellent recent reviews have discussed the complex mixture of advantages and disadvantages of polyploidy (see especially Spoelhof et al., 2017), and here we aimed to extend this with the most recent evidence considered explicitly in the scope of the dynamic temporal nature of shifting costs and benefits. In doing so, we hope to bring to light the importance of the timescales at which evolutionary dynamics are at play over the lifespan, from dawn till dusk, of any given genome duplication, thus creating the conditions for these wondrous cycles to emerge. We see a picture of each cycle of WGD-diploidization as a temporary but powerful engine of evolutionary diversification. Eventually, without specific selective pressures maintaining a strong advantage for polyploids, each hop to polyploidy is restabilized in a drop to diploid form, but there are plenty of evolutionary opportunities for speciation and radiation along the way.

AUTHOR CONTRIBUTIONS

PB drafted the manuscript and realized the figures with help from all other authors. SB drafted the manuscript with help from all other authors. MV-M drafted the manuscript with help from all other authors. FK drafted the manuscript with help from all other authors. LY drafted the manuscript with help from all other authors.

FUNDING

LY acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 679056) and the UK Biological and Biotechnology Research Council (BBSRC) via grant BB/P013511/1 to the John Innes Centre. Additional support was provided by Czech Science Foundation (project 17-20357Y to FK).

REFERENCES

Ågren, J. A., Huang, H.-R., and Wright, S. I. (2016). Transposable element evolution in the allotetraploid Capsella bursa-pastoris. *Am. J. Bot.* 103, 1197–1202. doi: 10.3732/ajb.1600103

Adams, K. L., and Wendel, J. F. (2005). Novel patterns of gene expression in polyploid plants. *Trends Genet.* 21, 539–543. doi: 10.1016/j.tig.2005.07.009

Agrawal, A. F., and Whitlock, M. C. (2011). Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics* 187, 553–566. doi: 10.1534/genetics.110.124560

Allario, T., Brumos, J., Colmenero-Flores, J. M., Iglesias, D. J., Pina, J. A., Navarro, L., et al. (2013). Tetraploid Rangpur lime rootstock increases drought tolerance via enhanced constitutive root abscisic acid production. *Plant Cell Environ.* 36, 856–868. doi: 10.1111/pce.12021

Anderson, J. K., and Warwick, S. I. (1999). Chromosome number evolution in the tribeBrassiceae (Brassicaceae): evidence from isozyme number. *Plant Syst. Evol.* 215, 255–285. doi: 10.1007/BF00984659

Arnold, B., Kim, S.-T., and Bomblies, K. (2015). Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Mol. Biol. Evol.* 32, 1382–1395. doi: 10.1093/molbev/msv089

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., et al. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci.* 113, 8320–8325. doi: 10.1073/pnas.1600405113

Arnold, M. L., and Kunte, K. (2017). Adaptive genetic exchange: a tangled history of admixture and evolutionary innovation. *Trends Ecol. Evol.* 32, 601–611. doi: 10.1016/J.TREE.2017.05.007

Arrigo, N., and Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* 15, 140–146. doi: 10.1016/j.pbi.2012.03.010

Baack, E. J. (2005). To succeed globally, disperse locally: effects of local pollen and seed dispersal on tetraploid establishment. *Heredity* 94, 538–546. doi: 10.1038/sj.hdy.6800656

Baack, E. J., and Stanton, M. L. (2005). Ecological factors influencing tetraploid speciation in snow buttercups (*ranunculus adoneus*): niche differentiation and tetraploid establishment. *Evolution* 59, 1936. doi: 10.1554/05-168.1

Baduel, P., Hunter, B., Yeola, S., and Bomblies, K. (2018). Genetic basis and evolution of rapid cycling in railway populations of tetraploid *Arabidopsis arenosa*. *PLoS Genet.* 14:e1007510. doi: 10.1371/journal.pgen.1007510

Baker, H. G. (1965). "Characteristics and modes of origin of weeds," in *The Genetics of Colonizing Species* (Asilomar, CA: Academic Press), 147–168. Available online at: http://www.cabdirect.org.ezp-prod1.hul.harvard.edu/abstracts/19661605504.html;jsessionid=0903F3BC5F41C10D0D2169FF9710AE2D?freeview=true (Accessed October 6, 2015).

Baker, H. G. (1967). Support for Baker's law–as a rule. *Evolution* 21, 853–856. doi: 10.2307/2406780

Bardil, A., Tayalé, A., and Parisod, C. (2015). Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. *Plant J.* 82, 621–631. doi: 10.1111/tpj.12837

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., et al. (2008). Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* 25, 2445–2455. doi: 10.1093/molbev/msn187

Barrett, S. C. H., Colautti, R. I., and Eckert, C. G. (2008). Plant reproductive systems and evolution during biological invasion. *Mol. Ecol.* 17, 373–383. doi: 10.1111/j.1365-294X.2007.03503.x

Barringer, B. C. (2007). Polyploidy and self-fertilization in flowering plants. *Am. J. Bot.* 94, 1527–1533. doi: 10.3732/ajb.94.9.1527

Beaulieu, J. M., Leitch, I. J., and Knight, C. A. (2007a). Genome size evolution in relation to leaf strategy and metabolic rates revisited. *Ann Bot.* 99, 495–505. doi: 10.1093/aob/mcl271

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A., and Knight, C. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.* 179, 975–986. doi: 10.1111/j.1469-8137.2008.02528.x

Beaulieu, J. M., Moles, A. T., Leitch, I. J., Bennett, M. D., Dickie, J. B., and Knight, C. A. (2007b). Correlated evolution of genome size and seed mass. *New Phytol.* 173, 422–437. doi: 10.1111/j.1469-8137.2006.01919.x

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Ann. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Bilinski, P., Albert, P. S., Berg, J. J., Birchler, J. A., Grote, M. N., Lorant, A., et al. (2018). Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. *PLOS Genet.* 14:e1007162. doi: 10.1371/journal.pgen.1007162

Birchler, J. A. (2014). Facts and artifacts in studies of gene expression in aneuploids and sex chromosomes. *Chromosoma* 123:459–469. doi: 10.1007/s00412-014-0478-5

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62. doi: 10.1111/j.1469-8137.2009.03087.x

Blair, A. C., and Wolfe, L. M. (2004). The evolution of an invasive plant: an experimental study with *Silene latifolia*. *Ecology* 85, 3035–3042. doi: 10.1890/04-0341

Bomblies, K., Higgins, J. D., and Yant, L. (2015). Tansley review Meiosis evolves: adaptation to external and internal environments. *New Phytol.* 208, 306–323. doi: 10.1111/nph.13499

Bomblies, K., Jones, G., Franklin, C., Zickler, D., and Kleckner, N. (2016). The challenge of evolving stable polyploidy: could an increase in "crossover interference distance" play a central role? *Chromosoma* 125, 287–300. doi: 10.1007/s00412-015-0571-4

Bomblies, K., and Madlung, A. (2014). Polyploidy in the arabidopsis genus. *Chromosome Res.* 22, 117–134. doi: 10.1007/s10577-014-9416-x

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Bretagnolle, F., and Thompson, J. D. (1995). Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. *New Phytol.* 129, 1–22. doi: 10.1111/j.1469-8137.1995.tb03005.x

Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A.-C., et al. (2004). Polyploidy in arctic plants. *Biol. J. Linn. Soc.* 82, 521–536. doi: 10.1111/j.1095-8312.2004.00337.x

Brown, A. H. D., and Marshall, D. S. (1981). "Evolutionary changes accompanying colonization in plants," in *Evolution Today: Proceedings of 2nd International Congress of Systematics and Evolutionary Biology*, 351–363.

Brownfield, L., and Köhler, C. (2011). Unreduced gamete formation in plants: mechanisms and prospects. *J. Exp. Bot.* 62, 1659–1668. doi: 10.1093/jxb/erq371

Bruggmann, R., Bharti, A. K., Gundlach, H., Lai, J., Young, S., Pontaroli, A. C., et al. (2006). Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* 16, 1241–51. doi: 10.1101/gr.5338906

Burleigh, J. G., Bansal, M. S., Wehe, A., and Eulenstein, O. (2008). "Locating multiple gene duplications through reconciled trees," in *Research in Computational Molecular Biology* (Berlin; Heidelberg: Springer), 273–284.

Burns, J. H. (2004). A comparison of invasive and non-invasive dayflowers (Commelinaceae) across experimental nutrient and water gradients. *Divers. Distributions* 10, 387–397. doi: 10.1111/j.1366-9516.2004.00105.x

Campbell, C. S., and Desai, A. (2013). What the hec is up with mouse oocyte meiosis? *Dev. Cell* 25, 3–4. doi: 10.1016/j.devcel.2013.03.018

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* 34, 247–278.

Ceccarelli, M., Santantonio, E., Marmottini, F., Amzallag, G. N., and Cionini, P. G. (2006). Chromosome endoreduplication as a factor of salt adaptation in Sorghum bicolor. *Protoplasma* 227, 113–118. doi: 10.1007/s00709-005-0144-0

Chao, D., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B., et al. (2013). Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science* 341, 658–659. doi: 10.1126/science.1240561

Chapman, M. A., and Abbott, R. J. (2010). Introgression of fitness genes across a ploidy barrier. *New Phytol.* 186, 63–71. doi: 10.1111/j.1469-8137.2009.03091.x

Charlesworth, B. (1976). Recombination modification in a fluctuating environment. *Genetics* 83, 181–195.

Charlesworth, B. (1992). Evolutionary rates in partially self-fertilizing species. *Am. Nat.* 140, 126–48. doi: 10.1086/285406

Charlesworth, D., and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Ann. Rev. Ecol. Syst.* 18, 237–268. doi: 10.1146/annurev.es.18.110187.001321

Chen, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Ann. Rev. Plant Biol.* 58, 377–406. doi: 10.1146/annurev.arplant.58.032806.103835

Chen, Z. J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays* 28, 240–252. doi: 10.1002/bies.20374

Chrtek, J., Herben, T., Rosenbaumová, R., Münzbergová, Z., Dočkalová, Z., Zahradníček, J., et al. (2017). Cytotype coexistence in the field cannot be explained by inter-cytotype hybridization alone: linking experiments and computer simulations in the sexual species Pilosella echioides (Asteraceae). *BMC Evol. Biol.* 17:87. doi: 10.1186/s12862-017-0934-y

Clark, J. W., and Donoghue, P. C. J. (2017). Constraining the timing of whole genome duplication in plant evolutionary history. *Proc. Biol. Sci.* 284, 20170912. doi: 10.1098/rspb.2017.0912

Coate, J. E., and Doyle, J. J. (2010). Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biol. Evol.* 2, 534–546. doi: 10.1093/gbe/evq038

Coate, J. E., and Doyle, J. J. (2015). Variation in transcriptome size: are we getting the message? *Chromosoma* 124, 27–43. doi: 10.1007/s00412-014-0496-3

Coate, J. E., Schlueter, J. A., Whaley, A. M., and Doyle, J. J. (2011). Comparative evolution of photosynthetic genes in response to polyploid and nonpolyploid duplication. *Plant Physiol.* 155, 2081–2095. doi: 10.1104/pp.105.073304

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Cookson, S. J., Radziejwoski, A., and Granier, C. (2006). Cell and leaf size plasticity in arabidopsis: what is the role of endoreduplication? *Plant Cell Environ.* 29, 1273–1283. doi: 10.1111/j.1365-3040.2006.01506.x

Coyne, J. A., and Orr, H. A. (2004). *Speciation*. Sunderland, MA. Sinauer Associates, Inc.

Daehler, C. C. (1998). Variation in self-fertility and the reproductive advantage of self-fertility for an invading plant (Spartina alterniflora). *Evol. Ecol.* 12, 553–568. doi: 10.1023/A:1006556709662

del Pozo, J. C., and Ramirez-Parra, E. (2014). Deciphering the molecular bases for drought tolerance in A rabidopsis autotetraploids. *Plant Cell Environ.* 37, 2722–2737. doi: 10.1111/pce.12344

Dodsworth, S., Chase, M. W., and Leitch, A. R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* 180, 1–5. doi: 10.1111/boj.12357

Drunen, W. E. Van, and Husband, B. C. (2018). Immediate vs. evolutionary consequences of polyploidy on clonal reproduction in an autopolyploid plant. *Ann. Bot.* 122, 195–205. doi: 10.1093/aob/mcy071

Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., et al. (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* 29, 2150–2167. doi: 10.1105/tpc.17.00010

Ehrendorfer, F. (1980). "Polyploidy and distribution," in *Polyploidy* (Boston, MA: Springer US), 45–60.

Elmaghrabi, A. M., Ochatt, S., Rogers, H. J., and Francis, D. (2013). Enhanced tolerance to salinity following cellular acclimation to increasing NaCl levels in Medicago truncatula. *Plant Cell Tissue Organ Cult.* 114, 61–70. doi: 10.1007/s11240-013-0306-2

Escudero, M., Martín-Bravo, S., Mayrose, I., Fernández-Mazuecos, M., Fiz-Palacios, O., Hipp, A. L., et al. (2014) Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. *PLoS ONE* 9:e85266. doi: 10.1371/journal.pone.0085266

Favarger, C. (1984). Cytogeography and biosystematics. *Plant Biosyst.* 453–476. doi: 10.1016/B978-0-12-295680-5.50033-0

Fawcett, J. A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5737–5742. doi: 10.1073/pnas.0900906106

Feldman, M. W., Christiansen, F. B., and Brooks, L. D. (1980). Evolution of recombination in a constant environment. *Proc. Natl. Acad. Sci. U.S.A.* 77, 4838–4841. doi: 10.1073/PNAS.77.8.4838

Feldman, M. W., Otto, S. P., and Christiansen, F. B. (1996). Population genetic perspective on the evolution of recombination. *Ann. Rev. Genet.* 30, 261–295. doi: 10.1146/annurev.genet.30.1.261

Fialová, M., and Duchoslav, M. (2014). Response to competition of bulbous geophyte Allium oleraceum differing in ploidy level. *Plant Biol.* 16, 186–196. doi: 10.1111/plb.12042

Franks, S. J., Sim, S., and Weis, A. E. (2007). Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1278–1282. doi: 10.1073/pnas.0608379104

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., and Mummenhoff, K. (2011). Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* 16, 108–116. doi: 10.1016/J.TPLANTS.2010.11.005

Freeling, M., Xu, J., Woodhouse, M., and Lisch, D. (2015). A solution to the c-value paradox and the function of junk DNA: The genome balance hypothesis. *Mol. Plant* 8, 899–910. doi: 10.1016/j.molp.2015.02.009

Galloway, L. F., Etterson, J. R., and Hamrick, J. L. (2003). Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, Campanula americana. *Heredity* 90, 308–315. doi: 10.1038/sj.hdy.6800242

Garbutt, K., and Bazzaz, F. A. (1983). Leaf demography, flower production and biomass of diploid and tetraploid populations of *phlox drummondii* hook. on a soil moisture gradient. *New Phytol.* 93, 129–141. doi: 10.1111/j.1469-8137.1983.tb02698.x

Gerstein, A. C., and Otto, S. P. (2009). Ploidy and the causes of genomic evolution. *J. Heredity* 100, 571–581. doi: 10.1093/jhered/esp057

Glennon, K. L., Ritchie, M. E., and Segraves, K. A. (2014). Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecol. Lett.* 17, 574–582. doi: 10.1111/ele.12259

Goldblatt, P. (1980). "Polyploidy in angiosperms: monocotyledons," in *Polyploidy* (Boston, MA: Springer US), 219–239.

Grant, V. (1981). *Plant Speciation*. New York, NY: Columbia University Press. xii, 563p.-illus., maps, chrom. nos.. En *2nd Edn*. Maps, Chromosome numbers. General (KR, 198300748).

Grant, V., (1963). *The Origin of Adaptations.* New York, NY: Columbia University Press.

Gregory, T. R. (2001). The bigger the C-value, the larger the cell: Genome size and red blood cell size in vertebrates. *Blood Cells Mol. Dis.* 27, 830–843. doi: 10.1006/bcmd.2001.0457

Gregory, T. R., and Mable, B. K. (2005). Chapter 8—polyploidy in animals. *Evol. Genome* 427–517. doi: 10.1016/B978-012301463-4/50010-3

Greiner, R., and Oberprieler, C. (2012). The role of inter-ploidy block for reproductive isolation of the diploid Leucanthemum pluriflorum Pau (Compositae, Anthemideae) and its tetra- and hexaploid relatives. *Flora Morphol. Distrib. Funct. Ecol. Plants* 207, 629–635. doi: 10.1016/J.FLORA.2012.07.001

Grelon, M., Vezon, D., Gendrot, G., and Pelletier, G. (2001). AtSPO11-1 is necessary for efficient meiotic recombination in plants. *EMBO J.* 20, 589–600. doi: 10.1093/emboj/20.3.589

Grotkopp, E., Rejmánek, M., and Rost, T. L. (2002). Toward a causal explanation of plant invasiveness: seedling growth and life-history strategies of 29 pine (Pinus) species. *Am. Nat.* 159, 396–419. doi: 10.1086/338995

Guo, M., Davis, D., and Birchler, J. A. (1996). Dosage effects on gene expression in a maize ploidy series. *Genetics* 142, 1349–1355.

Hagberg, A., and Ellerström, S. (1959). The competition between diploid, tetraploid and aneuploid rye: theoretical and practical aspects. *Hereditas* 45, 369–416. doi: 10.1111/j.1601-5223.1959.tb03058.x

Hagerup, O. (1932). Über polyploidie in Beziehung zu Klima, Ökologie und phylogenie. *Hereditas* 16, 19–40. doi: 10.1111/j.1601-5223.1932.tb02560.x

Hahn, M. A., Buckley, Y. M., and Müller-Schärer, H. (2012). Increased population growth rate in invasive polyploid *Centaurea stoebe* in a common garden. *Ecol. Lett.* 15, 947–954. doi: 10.1111/j.1461-0248.2012.01813.x

Haldane, J. B. S. (1932). *The Causes of Evolution*. Princeton, NJ: Princeton University Press.

Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *Am. Nat.* 67, 5–19. doi: 10.1086/280465

Haldane, J. B. S. (1937). The effect of variation of fitness. *Am. Nat.* 71, 337–349. doi: 10.1086/280722

Haldane, J. B. S., Harrison, J. W. H., Garrett, F. C., Gager, C. S., Blakeslee, A. F., and Metz, C. W. (1927). A mathematical theory of natural and artificial selection, part V: selection and mutation. *Math. Proc. Camb. Philos. Soc.* 23, 838. doi: 10.1017/S0305004100015644

Hall, M. C., and Willis, J. H. (2006). Divergent selection on flowering time contributes to local adaptation in Mimulus guttatus populations. *Evolution* 60, 2466–2477. doi: 10.1111/j.0014-3820.2006.tb01882.x

Hartung, F., Wurz-Wildersinn, R., Fuchs, J., Schubert, I., Suer, S. and Puchta, H. (2007). The catalytically active tyrosine residues of both SPO11-1 and SPO11-2 are required for meiotic double-strand break induction in *Arabidopsis*. *Plant Cell* 19, 3090–3099. doi: 10.1105/tpc.107.054817

Haufler, C. H. (1987). Electrophoresis is modifying our concepts of evolution in homosporous pteridophytes. *Am. J. Bot.* 74, 953. doi: 10.2307/2443877

Hegarty, M. J., Barker, G. L., Wilson, I. D., Abbott, R. J., Edwards, K. J., and Hiscock, S. J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. *Curr. Biol.* 16, 1652–1659. doi: 10.1016/J.CUB.2006.06.071

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Hill, R. R. (1971). Selection in autotetraploids. *Theoret. Appl. Genet.* 41, 181–186. doi: 10.1007/bf00277621

Hollingsworth, M. L., and Bailey, J. P. (2000). Evidence for massive clonal growth in the invasive weed Fallopia japonica (Japanese Knotweed). *Bot. J. Linn. Soc.* 133, 463–472. doi: 10.1006/bojl.2000.0359

Hollister, J. D. (2015). Polyploidy: adaptation to the genomic environment. *New Phytol.* 205, 1034–1039. doi: 10.1111/nph.12939

Hollister, J. D., Arnold, B. J., Svedin, E., Xue, K. S., Dilkes, B. P., and Bomblies, K. (2012). Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* 8:e1003093. doi: 10.1371/journal.pgen.1003093

Hollister, J. D., and Gaut, B. S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19, 1419–1428. doi: 10.1101/gr.091678.109

Horváth, V., Merenciano, M., and González, J. (2017). Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet.* 33, 832–841. doi: 10.1016/j.tig.2017.08.007

Hülber, K., Sonnleitner, M., Suda, J., Krejčíková, J., Schönswetter, P., Schneeweiss, G. M., et al. (2015). Ecological differentiation, lack of hybrids involving diploids, and asymmetric gene flow between polyploids in narrow contact zones of *Senecio carniolicus* (syn. *Jacobaea carniolica, Asteraceae*). *Ecol. Evol.* 5, 1224–1234. doi: 10.1002/ece3.1430

Husband, B. C. (2000). Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. *Proc. R. Soc. B Biol. Sci.* 267, 217–223. doi: 10.1098/rspb.2000.0990

Husband, B. C. (2004). The role of triploid hybrids in the evolutionary dynamics of mixed-ploidy populations. *Biol. J. Linn. Soc.* 82, 537–546. doi: 10.1111/j.1095-8312.2004.00339.x

Husband, B. C., Baldwin, S. J., and Sabara, H. A. (2016). Direct vs. indirect effects of whole-genome duplication on prezygotic isolation in *Chamerion angustifolium*: implications for rapid speciation. *Am. J. Bot.* 103, 1259–1271. doi: 10.3732/ajb.1600097

Husband, B. C., Ozimec, B., Martin, S. L., and Pollock, L. (2008). Mating consequences of polyploid evolution in flowering plants: current trends and insights from synthetic polyploids. *Int. J. Plant Sci.* 169, 195–206. doi: 10.1086/523367

Husband, B. C., and Sabara, H. A. (2003). Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, *Chamerion angustifolium* (Onagraceae). *New Phytol.* 161, 703–713. doi: 10.1046/j.1469-8137.2003.00998.x

Husband, B. C., and Schemske, D. W. (1997). The effect of inbreeding in diploid and tetraploid populations of *Epilobium angustifolium* (onagraceae): implications for the genetic basis of inbreeding depression. *Evolution* 51, 737–746. doi: 10.1111/j.1558-5646.1997.tb03657.x

Husband, B. C., Schemske, D. W., Burton, T. L., and Goodwillie, C. (2002). Pollen competition as a unilateral reproductive barrier between sympatric diploid and tetraploid *Chamerion angustifolium*. *Proc. R. Soc. B Biol. Sci.* 269, 2565–2571. doi: 10.1098/rspb.2002.2196

Ito, H., and Kakutani, T. (2014). Control of transposable elements in *Arabidopsis thaliana*. *Chromosome Res.* 22, 217–223. doi: 10.1007/s10577-014-9417-9

Ito, H., Kim, J. M., Matsunaga, W., Saze, H., Matsui, A., Endo, T. A., et al. (2016). A stress-activated transposon in arabidopsis induces transgenerational abscisic acid insensitivity. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep23181

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Johnston, M. O., and Schoen, D. J. (1996). Correlated evolution of self-fertilization and inbreeding depression: an experimental study of nine populations of Amsinckia (Boragainaceae). *Evolution* 50, 1478–1491.

Jørgensen, M. H., Ehrich, D., Schmickl, R., Koch, M. A., and Brysting, A. K. (2011). Interspecific and interploidal gene flow in Central European Arabidopsis (Brassicaceae). *BMC Evol. Biol.* 11:346. doi: 10.1186/1471-2148-11-346

Jung, C., and Müller, A. E. (2009). Flowering time control and applications in plant breeding. *Trends Plant Sci.* 14, 563–73. doi: 10.1016/j.tplants.2009.07.005

Kadereit, J. W. (2015). The geography of hybrid speciation in plants. *Taxon* 64, 673–687. doi: 10.12705/644.1

Kao, R. H. (2007). Asexuality and the coexistence of cytotypes. *New Phytol.* 175, 764–772. doi: 10.1111/j.1469-8137.2007.02145.x

Kashkush, K., Feldman, M., and Levy, A. A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* 33, 102–106. doi: 10.1038/ng1063

Knight, C. A., and Beaulieu, J. M. (2008). Genome size scaling through phenotype space. *Ann. Bot.* 101, 759–766. doi: 10.1093/aob/mcm321

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Köhler, C., Mittelsten Scheid, O., and Erilova, A. (2010). The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.* 26, 142–148. doi: 10.1016/J.TIG.2009.12.006

Kolár, F., Certner, M., Suda, J., Schönswetter, P., and Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Plant Sci.* 22, 1041–1055. doi: 10.1016/J.TPLANTS.2017.09.011

Korona, R. (1999). Unpredictable fitness transitions between haploid and diploid strains of the genetically loaded yeast *Saccharomyces cerevisiae*. *Genetics* 151, 77–85.

Koutecký, P., Badurová, T., Štech, M., Košnar, J., and Karásek, J. (2011). Hybridization between diploid Centaurea pseudophrygia and tetraploid *C. jacea* (Asteraceae): the role of mixed pollination, unreduced gametes, and mentor effects. *Biol. J. Linn. Soc.* 104, 93–106. doi: 10.1111/j.1095-8312.2011.01707.x

Kubátová, B., Trávníček, P., Bastlová, D., Curn, V., Jarolímová, V., and Suda, J. (2007). DNA ploidy-level variation in native and invasive populations of *Lythrum salicaria* at a large geographical scale. *J. Biogeogr.* 35, 167-176. doi: 10.1111/j.1365-2699.2007.01781.x

Lafon-Placette, C., Johannessen, I. M., Hornslien, K. S., Ali, M. F., Bjerkan, K. N., Bramsiepe, J., et al. (2017). Endosperm-based hybridization barriers explain the pattern of gene flow between Arabidopsis lyrata and *Arabidopsis arenosa* in Central Europe. *Proc. Natl. Acad. Sci. U.S.A.* 114, E1027–E1035. doi: 10.1073/pnas.1615123114

Lafuma, L., Balkwill, K., Imbert, E., Verlaque, R., and Maurice, S. (2003). Ploidy level and origin of the European invasive weed Senecio inaequidens (Asteraceae). *Plant Syst. Evol.* 243, 59–72. doi: 10.1007/s00606-003-0075-0

Lampson, M. A., and Cheeseman, I. M. (2011). Sensing centromere tension: aurora B and the regulation of kinetochore function. *Trends Cell Biol.* 21, 133–140. doi: 10.1016/J.TCB.2010.10.007

Lande, R., and Schemske, D. W. (1985). The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* 39, 24–40. doi: 10.2307/2408514

Le Comber, S. C., Ainouche, M. L., Kovarik, A., and Leitch, A. R. (2010). Making a functional diploid: from polysomic to disomic inheritance. *New Phytol.* 186, 113–122. doi: 10.1111/j.1469-8137.2009.03117.x

Leitch, A. R., and Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–3. doi: 10.1126/science.1153585

Leitch, I. J., and Bennett, M. D. (2004). Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* 82, 651–663. doi: 10.1111/j.1095-8312.2004.00349.x

Leitch, I. J., Hanson, L., Lim, K. Y., Kovarik, A., Chase, M. W., Clarkson, J. J., et al. (2008). The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). *Ann. Bot.* 101, 805–814. doi: 10.1093/aob/mcm326

Lenormand, T., and Otto, S. P. (2000). The evolution of recombination in a heterogeneous environment. *Genetics* 156, 423–438.

Levin, D. A. (1975). Minority cytotype exclusion in local plant populations. *Taxon* 24, 35. doi: 10.2307/1218997

Levin, D. A. (1983). Polyploidy and novelty in flowering plants. *Am. Nat.* 122, 3–147.

Levin, D. A. (2002). *The Role of Chromosomal Change in Plant Evolution*. New York, NY: Oxford University Press. Available online at: https://books.google.fr/books?hl=en&lr=&id=XinRCwAAQBAJ&oi=fnd&pg=PR9&dq=levin$+$2002$+$polyploidy&ots=yKoYFhdKcV&sig=nvaDxt6y7DE9mD6QXttyO6Ermao&redir_esc=y#v=onepage&q&f=false (Accessed April 14, 2018).

Lewontin, R. C. (1971). The effect of genetic linkage on the mean fitness of a population. *Proc. Natl. Acad. Sci. U.S.A.* 68, 984–986. doi: 10.1073/PNAS.68.5.984

Lim, K. Y., Kovarik, A., Matyasek, R., Chase, M. W., Clarkson, J. J., Grandbastien, M. A., et al. (2007). Sequence of events leading to near-complete genome turnover in allopolyploid Nicotiana within five million years. *New Phytol.* 175, 756–763. doi: 10.1111/j.1469-8137.2007.02121.x

Lopes, F. R., Jjingo, D., Da Silva, C. R. M., Andrade, A. C., Marraccini, P., Teixeira, J. B., et al. (2013). Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. *PLoS ONE* 8:e78931. doi: 10.1371/journal.pone.0078931

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., et al. (2012). Revisiting global gene expression analysis. *Cell* 151, 476–482. doi: 10.1016/J.CELL.2012.10.012

Luo, M. C., You, F. M., Li, P., Wang, J. R., Zhu, T., Dandekar, A. M., et al. (2015). Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *BMC Genomics* 16:707. doi: 10.1186/s12864-015-1906-5

Lysak, M. A., Cheung, K., Kitschke, M., and Bures, P. (2007). Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. *Plant Physiol.* 145, 402–410. doi: 10.1104/pp.107.104380

Mable, B. K., and Otto, S. P. (2001). Masking and purging mutations following EMS treatment in haploid, diploid and tetraploid yeast (*Saccharomyces cerevisiae*). *Genet. Res.* 77, 9–26.

Maceira, N. O., Jacquard, P., and Lumaret, R. (1993). Competition between diploid and derivative autotetraploid Dactylis glomerata L. from Galicia. Implications for the establishment of novel polyploid populations. *New Phytol.* 124, 321–328.

Madlung, A., Tyagi, A. P., Watson, B., Jiang, H., Kagochi, T., Doerge, R. W., et al. (2005). Genomic changes in synthetic Arabidopsis polyploids. *Plant J.* 41, 221–230. doi: 10.1111/j.1365-313X.2004.02297.x

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459. doi: 10.1073/pnas.0501102102

Mandák, B., Bímová, K., Pyšek, P., Štěpánek, J., and Plačková, I. (2005). Isoenzyme diversity in Reynoutria (Polygonaceae) taxa: escape from sterility by hybridization. *Plant Syst. Evol.* 253, 219–230. doi: 10.1007/s00606-005-0316-6

Mandák, B., Pyšek, P., Lysák, M., Suda, J., Krahulcová, A., and Bímová, K. (2003). Variation in DNA-ploidy levels of reynoutria taxa in the Czech Republic. *Ann. Bot.* 92, 265–272. doi: 10.1093/aob/mcg141

Mandáková, T., Li, Z., Barker, M. S., and Lysak, M. A. (2017). Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant J.* 91, 3–21. doi: 10.1111/tpj.13553

Mandáková, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42, 55–65. doi: 10.1016/j.pbi.2018.03.001

Manning, J. T., and Dickson, D. P. E. (1986). Asexual reproduction, polyploidy and optimal mutation rates. *J. Theoret. Biol.* 118, 485–489. doi: 10.1016/S0022-5193(86)80166-7

Masterson, J., (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264, 421–424. doi: 10.1126/science.264.5157.421

Masuyama, S., and Watano, Y. (1990). Trends for inbreeding in polyploid pteridophytes. *Plant Species Biol.* 5, 13–17.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Arrigo, N., Barker, M. S., Rieseberg, L. H., et al. (2015). Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytol.* 206, 27–35. doi: 10.1111/nph.13192

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., et al. (2011). Recently formed polyploid plants diversify at lower rates. *Science* 333, 1257. doi: 10.1126/science.1207205

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801. doi: 10.1126/science.15739260

Meng, H. B., Jiang, S. S., Hua, S. J., Lin, X. Y., Li, Y. L., Guo, W. L., et al. (2011). Comparison between a tetraploid turnip and its diploid progenitor (*Brassica rapa* L.): the adaptation to salinity stress. *Agric. Sci. China* 10, 363–375. doi: 10.1016/S1671-2927(11)60015-1

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends Genet.* 23, 183–191. doi: 10.1016/j.tig.2007.02.006

Moody, M. E., Mueller, L. D., and Soltis, D. E. (1993). Genetic variation and random drift in autotetraploid populations. *Genetics* 134, 649–657.

Mráz, P., Garcia-Jacas, N., Gex-Fabry, E., Susanna, A., Barres, L., and Müller-Schärer, H. (2012). Allopolyploid origin of highly invasive Centaurea stoebe s.l. (*Asteraceae*). *Mol. Phylogenet. Evol.* 62, 612–623. doi: 10.1016/J.YMPEV.2011.11.006

Müntzing, A. (1936). The evolutionary significance of autopolyploidy. *Hereditas* 21, 363–378. doi: 10.1111/j.1601-5223.1936.tb03204.x

Münzbergová, Z. (2007). Population dynamics of diploid and hexaploid populations of a perennial herb. *Ann. Bot.* 100, 1259–1270. doi: 10.1093/aob/mcm204

Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., et al. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* 20, 1545–57. doi: 10.1101/gr.109744.110

Novikova, P. Y., Hohmann, N., and Van de Peer, Y. (2018). Polyploid Arabidopsis species originated around recent glaciation maxima. *Curr. Opin. Plant Biol.* 42, 8–15. doi: 10.1016/j.pbi.2018.01.005

Osborn, T. C., Chris Pires, J., Birchler, J. A., Auger, D. L., Jeffery Chen, Z., Lee, H.-S., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* 19, 141–147. doi: 10.1016/S0168-9525(03)00015-5

Oswald, B. P., and Nuismer, S. L. (2011). A unified model of autopolyploid establishment and evolution. *Am. Nat.* 178, 687–700. doi: 10.1086/662673

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462. doi: 10.1016/j.cell.2007.10.022

Otto, S. P., and Michalakis, Y. (1998). The evolution of recombination in changing environments. *Trends Ecol. Evol.* 13, 145–151. doi: 10.1016/S0169-5347(97)01260-3

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* 34:401–437. doi: 10.1146/annurev.genet.34.1.40

Page, S. L., and Hawley, R. S. (2003). Chromosome chreography: the meiotic ballet. *Science* 301, 785–789. doi: 10.1126/science.1086605

Pandit, M. K., Pocock, M. J. O., and Kunin, W. E. (2011). Ploidy influences rarity and invasiveness in plants. *J. Ecol.* 99, 1108–1115. doi: 10.1111/j.1365-2745.2011.01838.x

Pandit, M. K., Tan, H. T. W., and Bisht, M. S. (2006). Polyploidy in invasive plant species of Singapore. *Bot. J. Linn. Soc.* 151, 395–403. doi: 10.1111/j.1095-8339.2006.00515.x

Pannell, J. R., and Barrett, S. C. H. (1998). Baker's law revisited: reproductive assurance in a metapopulation. *Evolution* 52, 657–668. doi: 10.1111/j.1558-5646.1998.tb03691.x

Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197. doi: 10.1038/nature 01771

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., and Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. *New Phytol.* 184, 1003–1015. doi: 10.1111/j.1469-8137.2009.0 3029.x

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9903–9908. doi: 10.1073/pnas.0307901101

Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C., and Doyle, J. J. (2005). Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* 54, 441–454. doi: 10.1080/10635150590945359

Piednoël, M., Sousa, A., and Renner, S. S. (2015). Transposable elements in a clade of three tetraploids and a diploid relative, focusing on Gypsy amplification. *Mob. DNA* 6, 5. doi: 10.1186/s13100-015-0034-8

Prentis, P. J., Wilson, J. R. U., Dormontt, E. E., Richardson, D. M., and Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends Plant Sci.* 13, 288–94. doi: 10.1016/j.tplants.2008.03.004

Price, S. C., and Jain, S. K. (1981). Are inbreeders better colonizers? *Oecologia* 49, 283–286. doi: 10.1007/BF00349202

Pyšek, P., Jarošík, V., Pergl, J., Randall, R., Chytrý, M., Kühn, I., et al. (2009). The global invasion success of Central European plants is related to distribution characteristics in their native range and species traits. *Divers. Distributions* 15, 891–903. doi: 10.1111/j.1472-4642.2009. 00602.x

Quadrana, L., Silveira, A. B., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., et al. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. *eLife* 5, 1–25. doi: 10.7554/eLife.15716

Ramsey, J. (2011). Polyploidy and ecological adaptation in wild yarrow. *Proc. Natl. Acad. Sci.* 108, 7096–7101. doi: 10.1073/pnas.1016631108

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Ramsey, J., and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* 33, 589–639. doi: 10.1146/annurev.ecolsys.33.010802.150437

Raven, P. H. (1975). The bases of angiosperm phylogeny: cytology. *Ann. Missouri Bot. Garden* 62, 724. doi: 10.2307/2395272

Renny-Byfield, S., Ainouche, M., Leitch, I. J., Lim, K. Y., Le Comber, S. C., and Leitch, A. R. (2010). Flow cytometry and GISH reveal mixed ploidy populations and Spartina nonaploids with genomes of S. *alterniflora* and *S. maritima* origin. *Ann. Bot.* 105, 527–533. doi: 10.1093/aob/mcq008

Rey, P. J., Manzaneda, A. J., and Alcántara, J. M. (2017). The interplay between aridity and competition determines colonization ability, exclusion and ecological segregation in the heteroploid Brachypodium distachyon species complex. *New Phytol.* 215, 85–96. doi: 10.1111/nph.14574

Richards, C. L., Bossdorf, O., Muth, N. Z., Gurevitch, J., and Pigliucci, M. (2006). Jack of all trades, master of some? On the role of phenotypic plasticity in plant invasions. *Ecol. Lett.* 9, 981–993. doi: 10.1111/j.1461-0248.2006.00950.x

Rodríguez, D. J. (1996). A model for the establishment of polyploidy in plants. *Am. Nat.* 147, 33–46. doi: 10.2307/2463222

Ronfort, J. (1999). The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. *Genet. Res.* 74, 31–42. doi: 10.1017/S0016672399003845

Roose, M. L., and Gottlieb, L. D. (1976). Genetic and biochemical consequences of polyploidy in tragopogon. *Evolution* 30, 818. doi: 10.2307/2407821

Ruprecht, C., Lohaus, R., Vanneste, K., Mutwil, M., Nikoloski, Z., Van De Peer, Y., and Persson, S. (2017). Revisiting ancestral polyploidy in plants. *Science Adv.* 3, 1–7. doi: 10.1126/sciadv.1603195

Sabara, H. A., Kron, P., and Husband, B. C. (2013). Cytotype coexistence leads to triploid hybrid production in a diploid-tetraploid contact zone of Chamerion angustifolium (Onagraceae). *Am. J. Bot.* 100, 962–970. doi: 10.3732/ajb.1200583

Saleh, B., Allario, T., Dambier, D., Ollitrault, P., and Morillon, R. (2008). Tetraploid citrus rootstocks are more tolerant to salt stress than diploid. *C. R. Biol.* 331, 703–710. doi: 10.1016/j.crvi.2008.06.007

Salman-Minkov, A., Sabath, N., and Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nat. Plants* 2, 16115. doi: 10.1038/nplants.2016.115

Sano, Y., Morishima, H., and Oka, H.-I. (1980). Intermediate perennial-annual populations ofOryza perennis found in Thailand and their evolutionary significance. *Bot. Mag. (Tokyo)* 93, 291–305. doi: 10.1007/BF02488735

Sarilar, V., Marmagne, A., Brabant, P., Joets, J., and Alix, K. (2011). BraSto, a Stowaway MITE from Brassica: recently active copies preferentially accumulate in the gene space. *Plant Mol. Biol.* 77, 59–75. doi: 10.1007/s11103-011-9794-9

Schemske, D. W., and Lande, R. (1985). The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. *Evolution* 39, 41. doi: 10.2307/2408515

Schinkel, C. C. F., Kirchheimer, B., Dellinger, A. S., Klatt, S., Winkler, M., Dullinger, S., et al. (2016). Correlations of polyploidy and apomixis with elevation and associated environmental gradients in an alpine plant. *AoB Plants* 8, plw064. doi: 10.1093/aobpla/plw064

Schlaepfer, D. R., Edwards, P. J., Semple, J. C., and Billeter, R. (2008). Cytogeography of *Solidago gigantea* (Asteraceae) and its invasive ploidy level. *J. Biogeogr.* 35, 2119–2127. doi: 10.1111/j.1365-2699.2008.01937.x

Schlueter, J. A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J. J., et al. (2004). Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47, 868–876. doi: 10.1139/g04-047

Schmickl, R., Marburger, S., Bray, S., and Yant, L. (2017). Hybrids and horizontal transfer: introgression allows adaptive allele discovery. *J. Exp. Bot.* 68, 5453–5470. doi: 10.1093/jxb/erx297

Schoustra, S. E., Debets, A. J. M., Slakhorst, M., and Hoekstra, R. F. (2007). Mitotic recombination accelerates adaptation in the fungus *Aspergillus nidulans*. *PLoS Genet.* 3:e68. doi: 10.1371/journal.pgen.0030068

Schranz, M. E., Mohammadin, S., and Edger, P. P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Curr. Opin. Plant Biol.* 15, 147–153. doi: 10.1016/J.PBI.2012.03.011

Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., et al. (2015) Polyploidy can drive rapid adaptation in yeast. *Nature* 519, 349–352. doi: 10.1038/nature14187

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

Sherrard, M. E., and Maherali, H. (2006). The adaptive significance of drought escape in Avena barbata, an annual grass. *Evolution* 60, 2478–2489. doi: 10.1111/j.0014-3820.2006.tb01883.x

Shi, T., Huang, H., and Barker, M. S. (2010). Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales. *Ann. Bot.* 106, 497–504. doi: 10.1093/aob/mcq129

Shi, X., Zhang, C., Ko, D. K., and Chen, Z. J. (2015). Genome-wide dosage-dependent and -independent regulation contributes to gene expression and evolutionary novelty in plant polyploids. *Mol. Biol. Evol.* 32, 2351–2366. doi: 10.1093/molbev/msv116

Simon-Porcar, V. I., Silva, J. L., Higgins, J. D., and Vallejo-Marin, M. (2017). Recent autopolyploidisation in a wild population of *Mimulus guttatus* (Phrymaceae). *Bot. J. Linn. Soc.* 185, 189–207. doi: 10.1093/botlinnean/box052

Slotte, T., Huang, H., Lascoux, M., and Ceplitis, A. (2008). Polyploid speciation did not confer instant reproductive isolation in Capsella (Brassicaceae). *Mol. Biol. Evol.* 25, 1472–1481. doi: 10.1093/molbev/msn092

Smith, H. E. (1946). Sedum pulchellum: a physiological and morphological comparison of diploid, tetraploid, and hexaploid races. *Bull. Torrey Bot. Club* 73, 495. doi: 10.2307/2481337

Solhaug, E. M., Ihinger, J., Jost, M., Gamboa, V., Marchant, B., Bradford, D., et al. (2016). Environmental regulation of heterosis in the allopolyploid *Arabidopsis suecica*. *Plant Physiol.* 170, 2251–2263. doi: 10.1104/pp.16.00052

Soltis, D. E., and Soltis, P. S. (1989). Genetic consequences of autopolyploidy in Tolmiea (Saxifragaceae). *Evolution* 43, 586. doi: 10.2307/2409061

Soltis, D. E., Soltis, P. S., and Rieseberg, L. H. (1993). Molecular data and the dynamic nature of polyploidy. *Crit. Rev. Plant Sci.* 12, 243–273. doi: 10.1080/07352689309701903

Soltis, D. E., Soltis, P. S., Schemske, D. W., Hancock, J. F., Thompson, J. N., Husband, B. C., et al. (2007). Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56, 13–30. doi: 10.2307/25065732

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–66. doi: 10.3732/ajb.1500501

Soltis, P. S., Marchant, D. B., Van de Peer, Y., and Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125. doi: 10.1016/J.GDE.2015.11.003

Soltis, P. S., and Soltis, D. E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad. Sci.* 97, 7051–7057. doi: 10.1073/pnas.97.13.7051

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

Sonnleitner, M., Weis, B., Flatscher, R., García, P. E., Suda, J., Krejčíková, J., et al. (2013) Parental ploidy strongly affects offspring fitness in heteroploid crosses among three cytotypes of autopolyploid *Jacobaea carniolica* (Asteraceae). *PLoS ONE* 8:e78959. doi: 10.1371/journal.pone.0078959

Spoelhof, J. P., Soltis, P. S., and Soltis, D. E. (2017). Pure polyploidy: closing the gaps in autopolyploid research. *J. Syst. Evol.* 55, 340–352. doi: 10.1111/jse.12253

Springer, N. M., Lisch, D., and Li, Q. (2016). Creating order from Chaos: epigenome dynamics in plants with complex genomes. *Plant Cell* 28, 314–325. doi: 10.1105/tpc.15.00911

Ståhlberg, D. (2009). Habitat differentiation, hybridization and gene flow patterns in mixed populations of diploid and autotetraploid *Dactylorhiza maculata* s.l. (*Orchidaceae*). *Evol. Ecol.* 23, 295–328. doi: 10.1007/s10682-007-9228-y

Stacey, N. J., Kuromori, T., Azumi, Y., Roberts, G., Breuer, C., Wada, T., et al. (2006). Arabidopsis SPO11-2 functions with SPO11-1 in meiotic recombination. *Plant J.* 48, 206–216. doi: 10.1111/j.1365-313X.2006.02867.x

Stebbins, G. (1971). *Chromosomal Evolution in Higher Plants*. London: Edward Arnold Ltd.

Stebbins, G. L. (1938). Cytological characteristics associated with the different growth habits in the dicotyledons. *Am. J. Bot.* 25, 189–198.

Stephens, A. D., Snider, C. E., Haase, J., Haggerty, R. A., Vasquez, P. A., Gregory Forest, M., et al. (2013). Individual pericentromeres display coordinated motion and stretching in the yeast spindle. *J. Cell Biol.* 203, 407–416. doi: 10.1083/jcb.201307104

Storchova, Z., and Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nat. Rev. Mol. Cell Biol.* 5, 45–54. doi: 10.1038/nrm1276

Storlazzi, A., Gargano, S., Ruprich-Robert, G., Falque, M., David, M., Kleckner, N., et al. (2010). Recombination proteins mediate meiotic spatial chromosome organization and pairing. *Cell* 141, 94–106. doi: 10.1016/j.cell.2010.02.041

Stupar, R. M., Bhaskar, P. B., Yandell, B. S., Rensink, W. A., Hart, A. L., Ouyang, S., et al. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics* 176, 2055–67. doi: 10.1534/genetics.107.074286

Sutherland, B. L., and Galloway, L. F. (2017). Postzygotic isolation varies by ploidy level within a polyploid complex. *New Phytol.* 213, 404–412. doi: 10.1111/nph.14116

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., et al. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* 207, 454–467. doi: 10.1111/nph.13491

te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubesova, M., et al. (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* 109, 19–45. doi: 10.1093/aob/mcr277

Theodoridis, S., Randin, C., Broennimann, O., Patsiou, T., and Conti, E. (2013). Divergent and narrower climatic niches characterize polyploid species of European primroses in *Primula* sect. *Aleuritia*. *J. Biogeogr.* 40, 1278–1289. doi: 10.1111/jbi.12085

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–46. doi: 10.1101/gr.4708406

Thompson, K. A., Husband, B. C., and Maherali, H. (2015). No influence of water limitation on the outcome of competition between diploid and tetraploid *Chamerion angustifolium* (Onagraceae). *J. Ecol.* 103, 733–741. doi: 10.1111/1365-2745.12384

Thompson, S. L., and Whitton, J. (2006). Patterns of recurrent evolution and geographic parthenogenesis within apomictic polyploid Easter daises (Townsendia hookeri). *Mol. Ecol.* 15, 3389–3400. doi: 10.1111/j.1365-294X.2006.03020.x

Tittel-Elmer, M., Bucher, E., Broger, L., Mathieu, O., Paszkowski, J., and Vaillant, I. (2010). Stress-induced activation of heterochromatic transcription. *PLoS Genet.* 6:e1001175. doi: 10.1371/journal.pgen.1001175

Treier, U. A., Broennimann, O., Normand, S., Guisan, A., Schaffner, U., Steinger, T., et al. (2009). Shift in cytotype frequency and niche space in the invasive plant *Centaurea maculosa*. *Ecology* 90, 1366–1377. doi: 10.1890/08-0420.1

Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., et al. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev.* 26, 705–13. doi: 10.1101/gad.183871.111

Vallejo-Marín, M., and Hiscock, S. J. (2016). Hybridization and hybrid speciation under global change. *New Phytol.* 211, 1170–1187. doi: 10.1111/nph.14004

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., et al. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534, 102–105. doi: 10.1038/nature17951

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

Vicient, C. M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharav, A., Nevo, E., et al. (1999). Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. *Plant Cell* 11, 1769–1784. doi: 10.1105/TPC.11.9.1769

Visger, C. J., Wong, G. K., Zhang, Y., Soltis, P. S., Soltis, D. E., Affiliation, A., et al. (2017). Divergent gene expression levels between diploid and autotetraploid. *bioRxiv*. doi: 10.1101/169367

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science* 290, 2114–2117. doi: 10.1126/science.290.5499.2114

Wagner, W. H. (1970). Biosystematics and evolutionary noise. *Taxon* 19, 146. doi: 10.2307/1217945

Wang, Z., Wang, M., Liu, L., and Meng, F. (2013). Physiological and proteomic responses of diploid and tetraploid black locust (*Robinia pseudoacacia* L.) subjected to salt stress. *Int. J. Mol. Sci.* 14, 20299–20325. doi: 10.3390/ijms141020299

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102, 1753–6. doi: 10.3732/ajb.1500320

Willis, J. H. (1999). Inbreeding load, average dominance and the mutation rate for mildly deleterious alleles in *Mimulus guttatus*. *Genetics* 153, 1885–1898.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–9. doi: 10.1073/pnas.0811575106

Yaakov, B., and Kashkush, K. (2012). Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Mol. Biol.* 80, 419–427. doi: 10.1007/s11103-012-9957-3

Yamauchi, A., Hosokawa, A., Nagata, H., and Shimoda, M. (2004). Triploid bridge and role of parthenogenesis in the evolution of autopolyploidy. *Am. Nat.* 164, 101–112. doi: 10.1086/421356

Yant, L., and Bomblies, K. (2017). Genomic studies of adaptive evolution in outcrossing Arabidopsis species. *Curr. Opin. Plant Biol.* 36, 9–14. doi: 10.1016/j.pbi.2016.11.018

Yant, L., Hollister, J. D., Wright, K. M., Arnold, B. J., Higgins, J. D., Franklin, F. C. H., et al. (2013). Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr. Biol.* 23, 2151–2156. doi: 10.1016/j.cub.2013.08.059

Yu, Z., Haberer, G., Matthes, M., Rattei, T., Mayer, K. F. X., Gierl, A., et al. (2010). Impact of natural genetic variation on the transcriptome of autotetraploid *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17809–17814. doi: 10.1073/pnas.1000852107

Zhang, J., Liu, Y., Xia, E.-H., Yao, Q.-Y., Liu, X.-D., and Gao, L.-Z. (2015). Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 112, E7022–E7029. doi: 10.1073/pnas.1515170112

Zhang, X., Deng, M., and Fan, G. (2014). Differential transcriptome analysis between *Paulownia fortunei* and its synthesized autopolyploid. *Int. J. Mol. Sci.* 15, 5079–5093. doi: 10.3390/ijms15035079

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., et al. (2016). Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Mol. Ecol.* 25, 2413–2426. doi: 10.1111/mec.13644

Zörgö, E., Chwialkowska, K., Gjuvsland, A. B., Garré, E., Sunnerhagen, P., Liti, G., et al. (2013). Ancient evolutionary trade-offs between yeast ploidy states. *PLoS Genet.* 9:e1003388. doi: 10.1371/journal.pgen.1003388

Conflict of Interest Statement: 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 30 April 2018
doi: 10.3389/fevo.2018.00052

Phylogenetic Structure of Plant Communities: Are Polyploids Distantly Related to Co-occurring Diploids?

Michelle L. Gaynor[1], *Julienne Ng*[2] *and Robert G. Laport*[2]*

[1] Department of Biology, University of Central Florida, Orlando, FL, United States, [2] Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO, United States

Polyploidy is widely acknowledged to have played an important role in the evolution and diversification of vascular plants. However, the influence of genome duplication on population-level dynamics and its cascading effects at the community level remain unclear. In part, this is due to persistent uncertainties over the extent of polyploid phenotypic variation, and the interactions between polyploids and co-occurring species, and highlights the need to integrate polyploid research at the population and community level. Here, we investigate how community-level patterns of phylogenetic relatedness might influence escape from minority cytotype exclusion, a classic population genetics hypothesis about polyploid establishment, and population-level species interactions. Focusing on two plant families in which polyploidy has evolved multiple times, Brassicaceae and Rosaceae, we build upon the hypothesis that the greater allelic and phenotypic diversity of polyploids allow them to successfully inhabit a different geographic range compared to their diploid progenitor and close relatives. Using a phylogenetic framework, we specifically test (1) whether polyploid species are more distantly related to diploids within the same community than co-occurring diploids are to one another, and (2) if polyploid species tend to exhibit greater ecological success than diploids, using species abundance in communities as an indicator of successful establishment. Overall, our results suggest that the effects of genome duplication on community structure are not clear-cut. We find that polyploid species tend to be more distantly related to co-occurring diploids than diploids are to each other. However, we do not find a consistent pattern of polyploid species being more abundant than diploid species, suggesting polyploids are not uniformly more ecologically successful than diploids. While polyploidy appears to have some important influences on species co-occurrence in Brassicaceae and Rosaceae communities, our study highlights the paucity of available geographically explicit data on intraspecific ploidal variation. The increased use of high-throughput methods to identify ploidal variation, such as flow cytometry and whole genome sequencing, will greatly aid our understanding of how such a widespread, radical genomic mutation influences the evolution of species and those around them.

Keywords: Brassicaceae, genome duplication, non-native species, phylogenetic community ecology, polyploidy, Rosaceae

OPEN ACCESS

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Jacob A. Tennessen,
Oregon State University, United States
Rubén Torices,
Université de Lausanne, Switzerland

****Correspondence:***
Robert G. Laport
rob.laport@gmail.com

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

Received: *22 January 2018*
Accepted: *11 April 2018*
Published: *30 April 2018*

Citation:
Gaynor ML, Ng J and Laport RG (2018) Phylogenetic Structure of Plant Communities: Are Polyploids Distantly Related to Co-occurring Diploids? Front. Ecol. Evol. 6:52. doi: 10.3389/fevo.2018.00052

INTRODUCTION

Polyploidy, or whole genome duplication, has been an important force shaping the evolutionary history of vascular plants (Adams and Wendel, 2005; Rieseberg and Willis, 2007; Soltis et al., 2009; Ramsey and Ramsey, 2014). Not only is polyploidy considered an important mechanism of speciation (Coyne and Orr, 2004; Soltis et al., 2014; Zhan et al., 2016), it is also often associated with major phenotypic shifts such as in size, flower color, water use, reproductive system, pollinator specialization, herbivore resistance, and phenology (Levin, 1983; Masterson, 1994; Segraves and Thompson, 1999; Husband et al., 2007; Maherali et al., 2009; Balao et al., 2011; Ramsey and Ramsey, 2014). Genome duplication has also been associated with novel alterations to genomic architecture and regulation that may affect adaptation (Comai, 2005; Madlung, 2013). However, despite the prevalence of polyploid events, the biodiversity implications of genome duplication, and the phenotypic differences often observed between diploids and polyploids, much remains unknown about how far reaching the impact of whole genome duplication is on interactions with other species and at the community level (Laport and Ng, 2017; Segraves, 2017).

Renewed interest in studying polyploidy over the last several decades has bent recent opinion toward acknowledging the significance of genome duplication on patterns of biodiversity (Coyne and Orr, 2004; Soltis et al., 2007; Ramsey and Ramsey, 2014; Laport and Ng, 2017; Segraves, 2017). Yet, the influence of genome duplication on population- and community-level dynamics remains unclear, in part because the evolutionary origin of polyploids may strongly influence the extent of polyploid phenotypic variation. Polyploids formed via the hybridization of two closely related species with partially diverged genomes (allopolyploidy) often exhibit phenotypes that are intermediate to, or outside the range of (i.e., transgressive), the parental species. In contrast, polyploids formed via the union of unreduced gametes within a population (autopolyploidy) often exhibit more subtle phenotypic differences when compared to their diploid progenitors. Historically, the more pronounced phenotypic variation among allopolyploids was considered as being important for interspecific interactions and patterns of biodiversity (Soltis et al., 2007; Ramsey and Ramsey, 2014). Research over the last few decades has shown, however, that the phenotypic, and underlying genetic, variation associated with both allo- and autopolyploids has the potential to influence ecological affinities, and play an important role in facilitating the establishment of new cytotypes, their expansion into a broader range of environmental conditions, and consequently their interactions with other species.

From the extensive body of empirical and theoretical work on the ecology and evolution of polyploids, whole genome duplication can be expected to have cascading effects on interspecific interactions and community-level dynamics (Ramsey and Ramsey, 2014; Čertner et al., 2017; Laport and Ng, 2017; Segraves, 2017). Although the direction and strength of the effect remains unclear, a number of predictions can be made about how species interactions and co-occurrence may be shaped by whole genome duplication based on previous species- and population-level work. For example, when considering first generation polyploids (i.e., tetraploids), it is thought that these neopolyploids must immediately compete with their co-occurring diploid progenitor upon formation while suffering a distinct frequency-dependent reproductive disadvantage. Because the relatively rare tetraploids are most likely to mate with more abundant diploids, this disadvantage, known as minority cytotype exclusion, arises from the lower fitness realized through the production of inviable or infertile triploid hybrid offspring (Hagberg and Ellerström, 1959; Levin, 1975). With few or no potential mates with which to reproduce, the neopolyploid is effectively "bred to death." However, even slight differences between polyploids and diploids, such as phenological shifts in the timing of reproduction, reproductive strategy (e.g., sexual vs. asexual), and ecological differences, may satisfy theoretical requirements for successful escape from minority cytotype exclusion (Husband, 2000). By easing direct ecological competition and promoting assortative mating, phenotypic differences may allow neopolyploids to persist within the range of their diploid progenitors. Present day communities would therefore reflect signatures of these historic events, whereby polyploids will often co-occur with their close diploid relatives.

Alternatively, polyploids may overcome minority cytotype exclusion by dispersing to new, unexploited habitats and maintaining the exclusion of their progenitors. Theoretical work predicts that the likelihood of neopolyploids becoming established at their site of origin is very low (Fowler and Levin, 2016), while dispersal and exploitation of novel habitat due to phenotypic differences either accompanying or arising rapidly after genome duplication greatly increases the probability of persistence (Lewis, 1962; Kay, 1969; Leitch and Leitch, 2008; Levin and Soltis, 2017). Indeed, the phenotypic differences associated with polyploidy may be great enough to facilitate the establishment of new cytotypes and their expansion into a broader range of ecological and environmental conditions. For example, polyploids have been documented to differ in ecological niche affinities and adaptive traits (Ramsey, 2011; McIntyre, 2012; Laport et al., 2013; Glennon et al., 2014; Marchant et al., 2016), experience morphological and physiological differences affecting phenological and physiological rates (Beaulieu et al., 2008; Manzaneda et al., 2012; Laport et al., 2016; Rey et al., 2017), have unique interactions with herbivores and pollinators (Thompson et al., 1997; Kennedy et al., 2006; Arvanitis et al., 2007; Halverson et al., 2008; Thompson and Merg, 2008; Roccaforte et al., 2015), and exhibit unique water relations (Maherali et al., 2009) and mycorrhizal associations (Těšitelová et al., 2013). Shifts from sexual to asexual reproduction, or a breakdown of self-incompatibility systems (Comai, 2005; Otto, 2007), could further promote the establishment of polyploids in geographic areas isolated from their diploid progenitors by providing a means of reproduction and population increase. If strong ecological differentiation between cytotypes and establishment in geographically isolated areas is the predominant mode of neopolyploid success and persistence, polyploids should more often occur in different communities than their close diploid relatives.

In addition to phenotypic differences between polyploids and diploids, variation at the molecular level also likely bears strongly on community assembly. In particular, genetic changes associated with whole genome duplication could increase the ecological success of polyploids in novel communities. Doubled nuclear DNA content on its own can have cellular phenotypic consequences that alter intracellular stoichiometric relationships and physiological rates, causing shifts to growth rate, gas exchange, and flowering time (Comai, 2005; Beaulieu et al., 2008; Madlung, 2013; Bilinski et al., 2018), which may allow polyploids to outcompete co-occurring diploids. The novel genetic architecture and regulatory environment of duplicated genomes may also lead to greater adaptability, and the larger genome size may be a larger target for functional mutations that could influence adaptation (Comai, 2005; Madlung, 2013; Soltis et al., 2015; Song and Chen, 2015; Mei et al., 2018). For example, the increased genomic content of polyploids presents potential opportunities for rapid paralog subfunctionalization or neofunctionalization that could lead to greater competitive ability or ecological success, and even invasiveness, relative to diploid progenitors (Thompson and Lumaret, 1992; Schlaepfer et al., 2010; te Beest et al., 2011; Green et al., 2013; Pyšek et al., 2013; Nagy et al., 2017). Indeed, the increased genomic/allelic diversity of larger genomes, decreased inbreeding depression, multisomic inheritance, intergenomic recombination, and accelerated epigenetic processes of polyploids have been identified as major factors that may predispose polyploid populations to rapidly exploit novel ecological niches (Comai, 2005; Soltis et al., 2009; Parisod et al., 2010; Green et al., 2013; Madlung, 2013). Thus, while the genetic changes associated with whole genome duplication and their influence over ecologically relevant phenotypic shifts may be used as a basis to make predictions about the ecological success of polyploids within communities, it remains relatively unexplored whether polyploids are indeed better competitors in a community context.

One way to investigate the influence of genome duplication on community structure is by analyzing diploid and polyploid co-occurrence within multiple communities using a comparative phylogenetic framework. Although there has been an increase in studies integrating phylogenetic data with questions about community ecology over the last decade (Webb et al., 2002; Cavender-Bares et al., 2006; Emerson and Gillespie, 2008; Vamosi et al., 2009), no studies have explicitly included ploidal information to assess the influence of genome duplication (and associated phenotypes) on community structure. Here, we use a novel approach to examine how polyploids influence phylogenetic community structure by combining ploidal information with phylogenetic analyses of plant communities across the United States. Specifically, we focus on two large plant families that are well represented across North American biomes and in which polyploidy has evolved multiple times, Brassicaceae and Rosaceae, to test (1) whether polyploid species are more distantly related to diploids within the same community than co-occurring diploids are to one another. We expect this phylogenetic pattern if polyploids escaped minority cytotype exclusion by inhabiting a different geographic range compared to their diploid progenitor and close relatives. We also test (2) whether polyploid species tend to exhibit greater ecological success than diploid species, using the relative abundance of polyploids vs. diploids as an indicator of successful establishment within communities. We further compare the abundance of native and non-native species to examine whether species experiencing recent ecological range expansions (i.e., non-native species) also tend to be polyploid.

MATERIALS AND METHODS

Community Data Collection

We obtained species composition and abundance data for Brassicaceae and Rosaceae communities across the United States from the National Ecological Observatory Network (NEON; https://www.neonscience.org; Keller et al., 2008). NEON has established sites across the United States and conducted plant surveys of replicated 400 m^2 plots across each site. We specifically focused on Brassicaceae and Rosaceae communities because they are polyploid-rich, broadly represent contrasting life histories, and were present in a large number of NEON communities. We focused on 16 communities (**Figure 1**), each of which had three or more representatives from the respective family for which we could obtain ploidal data (6 Brassicaceae communities, 11 Rosaceae communities; **Figure 2**). For each species, we determined its ploidal level based on scientific literature and online databases (Kew *C*-value database, http://data.kew.org/cvalues/; Chromosome Count Database, https://ccdb.tau.ac.il; Table S1), as well as its native status following the designation assigned in the United States Department of Agriculture (USDA) PLANTS database (https://plants.usda.gov) (**Figure 2**). While mode of polyploid origin is likely important for interspecific interactions and ecological success, we were unable to consider differences in origin for this study as we could not consistently determine whether a species was an allo- or autopolyploid. As geographic variation in ploidy can be common (Baack, 2005; Kolár et al., 2009; Ståhlberg, 2009; Trávníček et al., 2011; Castro et al., 2012; Laport et al., 2012; Ramsey and Ramsey, 2014; Zozomová-Lihová et al., 2015; Wefferling et al., 2017), we aimed to determine the community-specific ploidal level of each species. When species were reported to comprise multiple ploidal levels for the region around a NEON site (**Figure 2**), we repeated analyses with each ploidy. When assigning native status to each species, we considered species to be either native or non-native to the lower 48 states.

Phylogenetic Reconstruction

As published phylogenies of Brassicaceae and Rosaceae did not include all members of our study communities (Huang et al., 2016; Zhang et al., 2017), we reconstructed phylogenies for each family using sequence data from GenBank and newly generated sequence data for species that did not have publicly available sequence data for our target genetic loci. We focused on one nuclear locus, ITS (internal transcribed spacer), and two chloroplast loci, rbcL (ribulose bisphosphate carboxylase large chain) and matK (maturase K). To generate our own sequences, leaf tissue was obtained from the Rocky Mountain Herbarium (RM) and the Missouri Botanical Gardens Herbarium (MO).

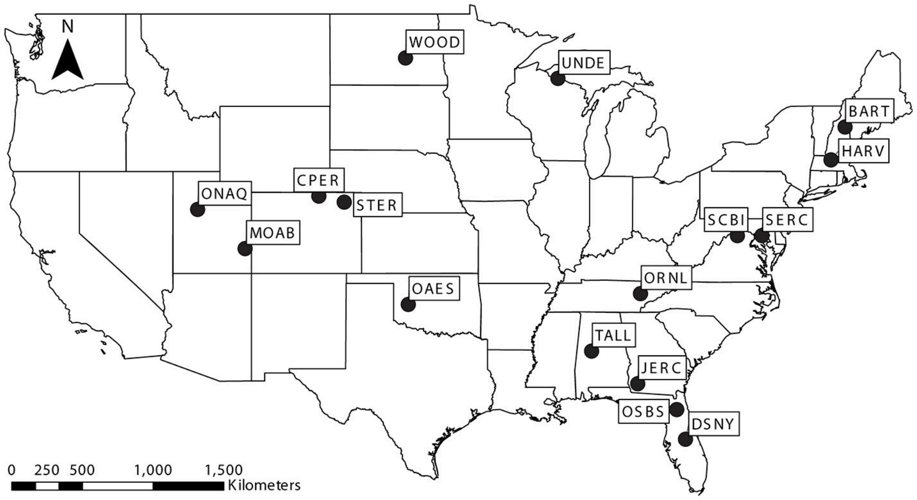


FIGURE 1 | Map of the United States showing study communities collected from established National Ecological Observatory Network sites: Bartlett Experimental Forest (BART), Central Plains Experimental Range (CPER), Disney Wilderness Preserve (DSNY), Harvard Forest (HARV), Jones Ecological Research Center (JERC), Moab (MOAB), Klemme Range Research Station (OAES), Onaqui-Ault (ONAQ), Oak Ridge National Laboratory (ORNL), Ordway-Swisher Biological Station (OSBS), Smithsonian Environmental Research Center (SERC), Smithsonian Conservation Biology Institute (SCBI), North Sterling (STER), Talladega National Forest (TALL), Woodworth (WOOD), and University of Notre Dame Environmental Research Center (UNDE).

DNA was extracted using the Qiagen Plant Mini Kit or CTAB DNA extraction method (Doyle and Doyle, 1987). We amplified the gene regions using previously published primers (Table S2) and following PCR protocols available in the Supplementary Materials (Supplementary Material 1). As we had difficulty amplifying ITS for Rosaceae due to polymorphisms in binding sites, we designed a new primer using Primer 3 (Koressaar and Remm, 2007; Untergasser et al., 2012) based on previously sequenced Rosaceae species: ITS_SGR (5′-AGG TTT GAC AAC CAC CGA TT-3′). We sent PCR products to Genewiz (Cambridge, Massachusetts) for purification and sequencing, and checked sequence quality in Geneious v6.0.5 (Biomatters Ltd., Auckland, NZ).

To ensure that the evolutionary relationships among members of the community were consistent with known relationships, we reconstructed phylogenies that included all species in this study, as well as any other available sequences from GenBank for each family. The inclusion of additional species not occurring within the communities of focus in phylogenetic reconstruction has been shown to reduce error in node age estimates, and consequently in calculations of community phylogenetic diversity metrics (Park et al., 2018). High quality sequence data for the targeted genetic loci were downloaded from GenBank using the PHLAWD pipeline (Smith et al., 2009). We combined GenBank sequences with newly generated sequences, aligned them in Mafft v7 (Katoh et al., 2002) and concatenated the gene regions in Mesquite v3.10 (Maddison and Maddison, 2017). The final data set included 1,912 species for Brassicaceae including five outgroup members (*Cleome lutea* Hook., *Cleome viscosa* L., *Cleome rutidosperma* DC., *Moringa oleifera* Lam., *Polanisia dodecandra* (L.) DC.). For Rosaceae, the final data set included 1,450 species including four outgroup members (*Rhamnus cathartica* L., *Ceanothus verrucosus* Nutt., *Pisum sativum* L., *Astragalus membranaceus* (Fisch.) Bunge).

We used Bayesian inference to reconstruct a time-calibrated phylogeny for each family using BEAST2 v2.4.5 (Bouckaert et al., 2014) on the CIPRES Science Gateway (www.phylo.org). For the phylogenetic reconstruction of Brassicaceae, the stem and crown nodes were constrained with a lognormal offset of 59.5 and 42.0 million years ago (Ma) (mean 0.01, standard deviation 1.0), respectively, following Huang et al. (2016). For the phylogenetic reconstruction of Rosaceae, the stem and crown nodes were constrained with a lognormal offset of 106.50 and 95.09 Ma (mean 0.01, standard deviation 1.0), respectively, following Zhang et al. (2017). We conducted two runs of 120 million generations and sampled trees every 12,000 generations. We used Tracer v1.6 (Rambaut et al., 2014) to verify that both runs reached stationarity and converged on the posterior distribution of trees. As identified in Tracer, we discarded 10% of the trees from each run as burn-in, then combined and summarized trees as a maximum clade credibility (MCC) tree using LogCombiner and TreeAnnotator (included as part of the BEAST2 package). We pruned all species that were not included in each of our study communities from the trees prior to site-specific analyses.

Diploid and Polyploid Phylogenetic Relationships

We used two approaches to determine whether polyploid species are more distantly related to diploids within the same community than co-occurring diploids are to one another. First, we used a broad-scale approach to investigate patterns of phylogenetic relatedness across all sites by calculating the phylogenetic distance between diploids and their closest diploid relative within the same community (nearest taxon distance; NTD_{2x-2x}), and comparing these distances to the phylogenetic distance between

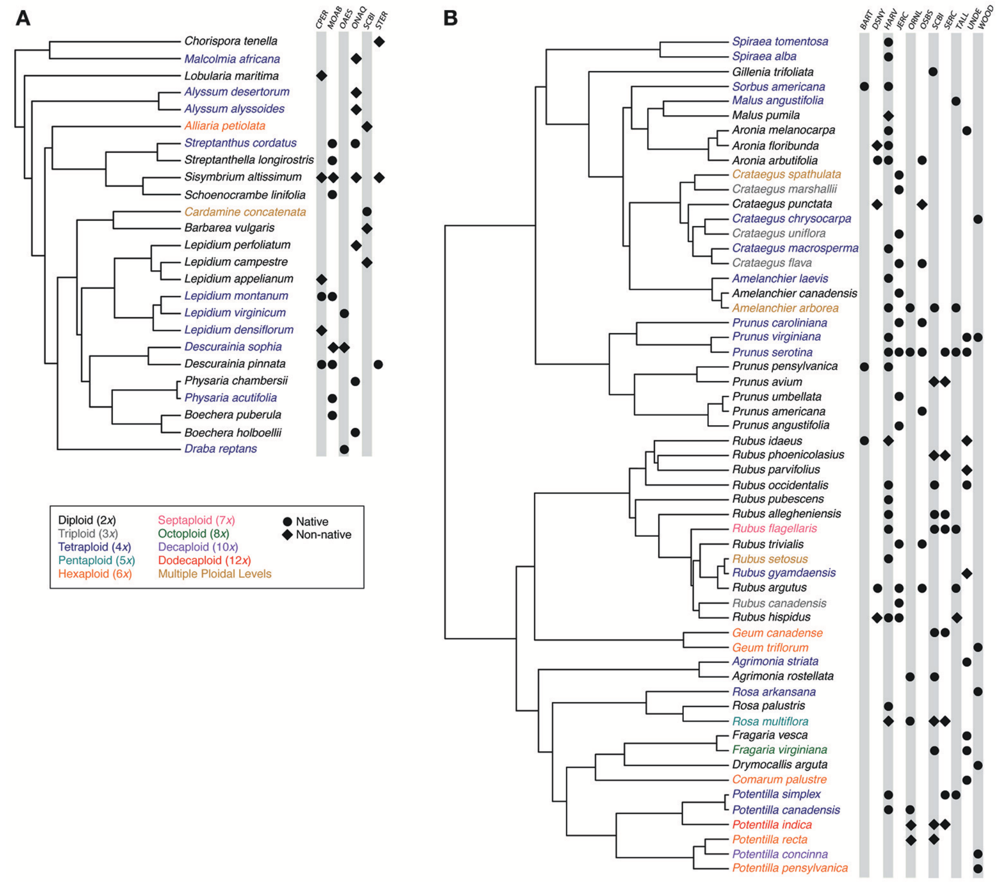


FIGURE 2 | Phylogenetic trees showing **(A)** Brassicaceae and **(B)** Rosaceae community members and their occurrence in each study community. Species names are colored to indicate ploidal level. Symbols indicate the species' presence within a study community, with circles and diamonds indicating whether the species is native or non-native, respectively. Site abbreviations follow **Figure 1**.

polyploids and their most closely related, co-occurring diploid species ($NTD_{polyploid-2x}$). We pooled these values across sites and compared NTD_{2x-2x} to $NTD_{polyploid-2x}$ by conducting a Mann-Whitney U test using the wilcox.test function in R. We also evaluated our hypothesis of closer relationships between co-occurring diploids than among co-occurring diploids and polyploids by comparing the proportion of NTD_{2x-2x} and $NTD_{polyploid-2x}$ comparisons that fell below a threshold of the mean nearest taxon distance (MNTD) of the family-level phylogeny. MNTD was calculated using the cophenetic.phylo function in the ape R package (Paradis et al., 2004).

Second, we examined patterns of phylogenetic relatedness within each site by testing whether the MNTD between polyploids and diploids ($MNTD_{polyploid-2x}$) was significantly greater than $MNTD_{2x-2x}$ than expected by chance. We employed a simulation approach by comparing the observed metric $MNTD_{polyploid-2x}$ / $MNTD_{2x-2x}$ within each community to a null distribution generated by replacing polyploid community members with randomly drawn species from a pool of polyploids from all sites, and recalculating the $MNTD_{polyploid-2x}$ / $MNTD_{2x-2x}$ metric for the new community. Our null distribution comprised 1,000 random communities per site. We considered polyploids to be more distantly related to diploids than expected by chance if the $MNTD_{polyploid-2x}$ / $MNTD_{2x-2x}$ metric was greater than 1 and was greater than 95% of the null distribution ($P < 0.05$). Any communities that did not have both diploid and polyploid species (Rosaceae: DSNY; Brassicaceae: STER, OAES) or only had one diploid or polyploid representative (Rosaceae: WOOD) were excluded from these analyses. All MNTD calculations were performed using the ses.mntd function in the picante R package (Kembel et al., 2010).

Tests of Polyploid Ecological Success

We identified whether polyploid species showed patterns consistent with having greater ecological success than diploids by using species abundance as an indicator of successful establishment within a community (Levin, 1975; Callaway and Aschehoug, 2000; Cleland et al., 2004). Specifically, we tested whether polyploids occurred at greater total relative abundance than diploids within each community by conducting a Mann-Whitney *U*-test with the wilcox.test function in R. We further assessed whether differences in abundance could be attributed to non-native species, reflecting ecological success of recent range expansions, by testing whether the total relative abundance of diploid and polyploid species significantly differed between natives and non-natives. We tested significance using a Kruskal-Wallis test and when appropriate, followed the analysis with Dunn's *post-hoc* test. This was performed using the kruskal.test and the dunnTest (FSA package) functions, respectively, in R.

RESULTS

Community Data Collection

Our six Brassicaceae communities comprised 3–8 Brassicaceae species while our eleven Rosaceae communities comprised 3–24 Rosaceae species (**Figure 2**). Most Brassicaceae in our communities were 2*x*, 4*x*, or 6*x*, with the exception of one species where the ploidy ranged from 20*x* to 30*x* [*Cardamine concatenata* (Michx.) O. Schwarz.; Kreiner et al., 2017; Table S1]. In Brassicaceae communities, 33–100% of the species were polyploid, and 22–75% of the species were non-native (**Figure 2**). In Rosaceae communities, species ranged in ploidal level from 2*x* to 12*x*, with 33–86% of the species being polyploid. These communities also ranged from not having any non-native species to 44% of the species being non-native.

Phylogenetic Reconstruction

We generated 63 new sequences for species missing sequence data for our target loci (GenBank accessions KY427264-KY427326; Table S3). The final Brassicaceae alignment comprised 1,912 species and was 8,242 basepairs (bp) in length, while the final Rosaceae alignment comprised 1,450 species and was 12,007 bp in length (TreeBASE accession: S22405). All study species within the communities were represented in our time-calibrated phylogenetic trees, and both phylogenies for Brassicaceae and Rosaceae community members were congruent in topology to previously published phylogenies (Huang et al., 2016; Zhang et al., 2017; **Figure 2**).

Diploid and Polyploid Phylogenetic Relationships

Our broad-scale analysis examining phylogenetic patterns of relatedness between co-occurring polyploids and diploids vs. co-occurring diploids found that across all sites, $\mathrm{NTD}_{\mathrm{polyploid}-2x}$ was significantly greater than NTD_{2x-2x} for both Brassicaceae ($P < 0.05$) and Rosaceae ($P << 0.01$; **Figure 3**). Further supporting this result for both families was that a larger proportion of $\mathrm{NTD}_{2\mathrm{x}-2\mathrm{x}}$ comparisons fell below the MNTD threshold compared to $\mathrm{NTD}_{\mathrm{polyploid}-2\mathrm{x}}$. For Brassicaceae communities, 76.9% of diploid-diploid comparisons and 34.1% of polyploid-diploid comparisons fell below the Brassicaceae MNTD, while for Rosaceae, 84.1% of diploid-diploid comparisons and 53.8% of polyploid-diploid comparisons fell below the Rosaceae MNTD (**Figure 3**). This pattern suggests that fewer polyploids co-occur with a close diploid relative compared to diploids.

When examining each site, $\mathrm{MNTD}_{\mathrm{polyploid}-2x}$ was greater than MNTD_{2x-2x} for three of the four Brassicaceae communities ($\mathrm{MNTD}_{\mathrm{polyploid}-2x}$ / $\mathrm{MNTD}_{2x-2x} > 1$; **Figure 4**). However, $\mathrm{MNTD}_{\mathrm{polyploid}-2x}$ / MNTD_{2x-2x} was only significantly greater than expected by chance at one site (ONAQ; $P < 0.05$) in our simulation analyses. At MOAB, although $\mathrm{MNTD}_{\mathrm{polyploid}-2x}$ was greater than MNTD_{2x-2x}, the phylogenetic distance was smaller than expected by chance (lower 2.5% of the null distribution).

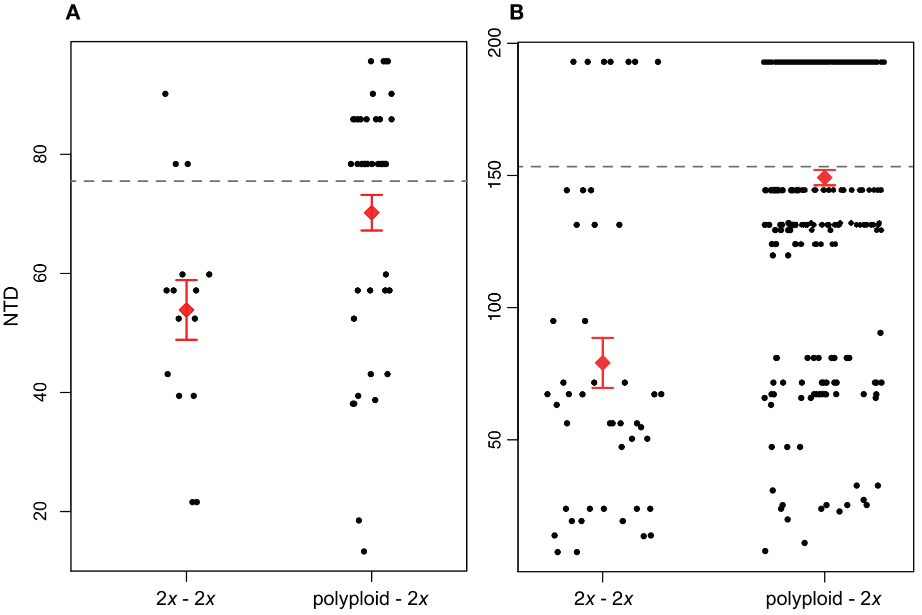


FIGURE 3 | Phylogenetic distance between diploids and their closest diploid relative (nearest taxon distance; NTD) within the same community (2*x* – 2*x*) and NTD between polyploids and diploids within the same community (polyploid – 2*x*). NTD differences between the two groups are significant for both **(A)** Brassicaceae ($P < 0.05$) and **(B)** Rosaceae ($P << 0.01$). The red diamond and error bars show the mean of the distribution ±1 standard error. The mean NTD (MNTD) for each family-level phylogeny is indicated by the dashed horizontal line.

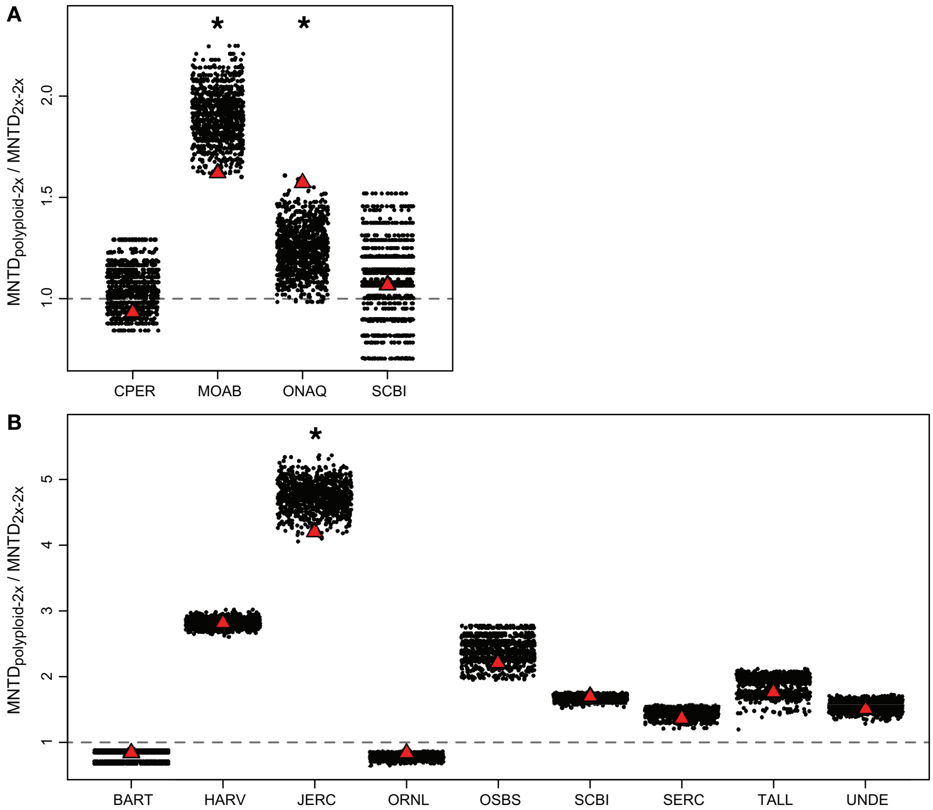


FIGURE 4 | Comparison of observed $\text{MNTD}_{\text{polyploid}-2x}$ / MNTD_{2x-2x} at each site (red triangle) to simulated random communities (black dots) for **(A)** Brassicaceae and **(B)** Rosaceae. The random expectation was generated by randomly replacing polyploid species from a pool of polyploids from all study communities. $\text{MNTD}_{\text{polyploid}-2x}$ / $\text{MNTD}_{2x-2x} > 1$ (above dashed line) indicates that the mean phylogenetic distance between polyploids and the closest diploid relative in the same community is greater than that between co-occurring diploids. Asterisks indicate that the observed $\text{MNTD}_{\text{polyploid}-2x}$ / MNTD_{2x-2x} is significantly different from random ($P < 0.05$). If the observed value is significantly higher than the random distribution, $\text{MNTD}_{\text{polyploid}-2x}$ / MNTD_{2x-2x} is significantly greater than expected by chance. Alternatively, if the observed value is significantly lower than the random distribution, $\text{MNTD}_{\text{polyploid}-2x}$ / MNTD_{2x-2x} is significantly less than expected by chance. The significant difference at JERC for Rosaceae **(B)** reflects the analysis with *Crataegus spathulata* as a diploid (vs. triploid). Site name abbreviations follow **Figure 1**.

Within Rosaceae communities, $\text{MNTD}_{\text{polyploid}-2x}$ was greater than MNTD_{2x-2x} for seven of the nine communities, but none of these differences were significantly different from the random expectation in our simulation analyses (**Figure 4**). For one community (JERC), we considered *Crataegus spathulata* Michx. to either be a diploid or a triploid. When analyzed as a diploid, we found that $\text{MNTD}_{\text{polyploid}-2x}$ was smaller than expected by chance, although overall, $\text{MNTD}_{\text{polyploid}-2x}$ was still greater than MNTD_{2x-2x} ($\text{MNTD}_{\text{polyploid}-2x}$ / $\text{MNTD}_{2x-2x} > 1$). However, we did not find any significant patterns when *C. spathulata* was treated as a triploid in the analyses. At another community (HARV), we performed analyses with *Rubus setosus* Bigelow as either diploid or triploid, however there was no effect on the overall results.

Tests of Polyploid Ecological Success

In Brassicaceae communities, polyploids tended to be more abundant than diploids ($P > 0.05$; **Figure 5A**). Though not a significant pattern, in all communities that included both diploid and polyploid species, ≥70% of the individuals were polyploid. The greater abundance of polyploids appears to be driven by non-native polyploids, which tended to be greater in number than native polyploids (**Figure 5C**). However, we did not find a significant difference in the abundance of non-native and native diploid or polyploid individuals within any of the communities ($P > 0.05$). This may be due to the small number of Brassicaceae communities included in the analysis, or could suggest that the ecological success of polyploid species is not the result of non-native species experiencing recent range expansions.

In Rosaceae communities, we found no significant difference between diploid and polyploid abundance ($P > 0.05$; **Figure 5B**). When polyploids and diploids were categorized as native or non-native, however, we found that native species were significantly more abundant than co-occurring non-native species for both diploids and polyploids ($P < 0.05$; **Figure 5D**) suggesting that ecological success is not necessarily associated with genome duplication.

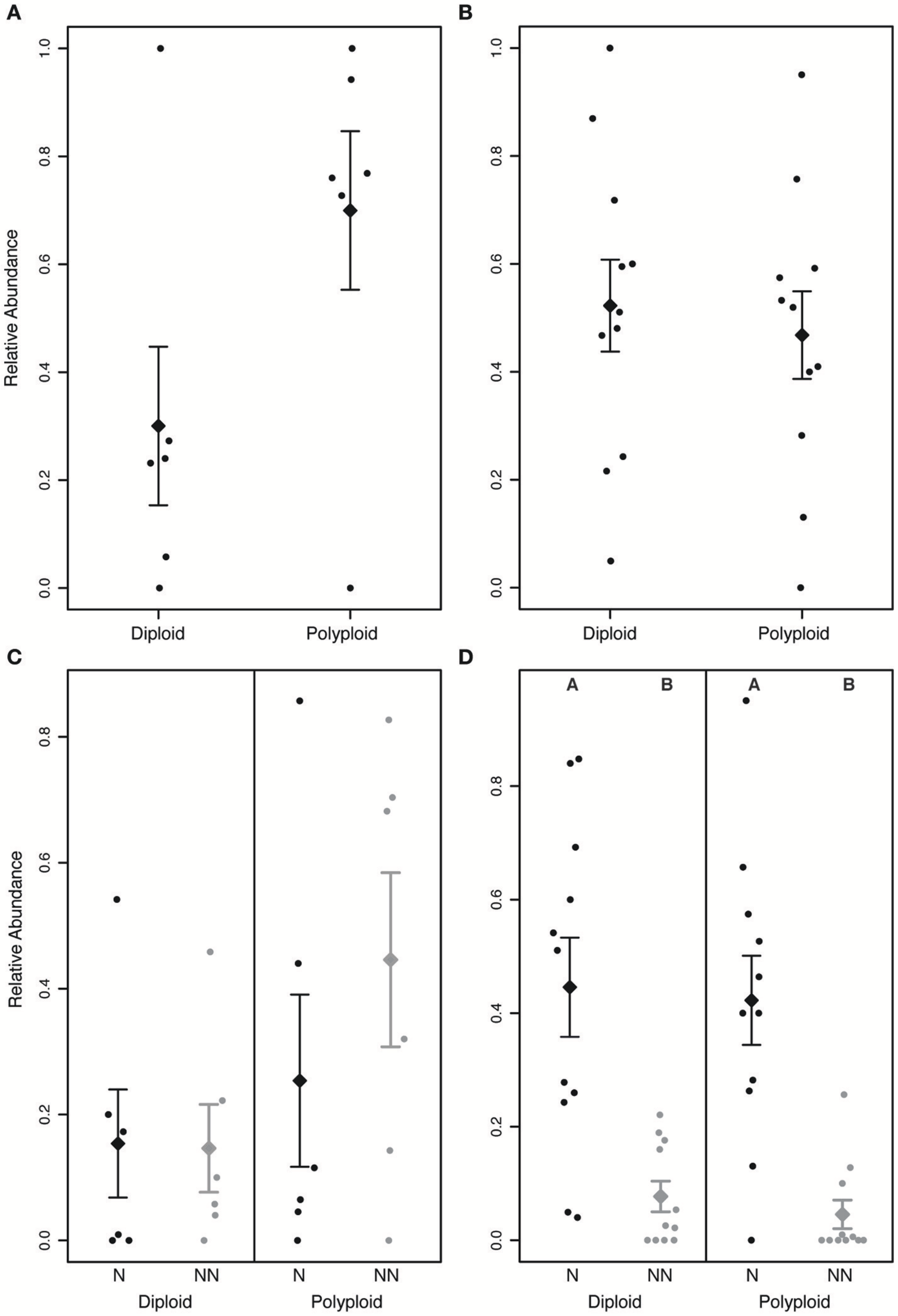


FIGURE 5 | Relative abundance of diploids and polyploids in **(A,C)** Brassicaceae and **(B,D)** Rosaceae communities. **(A)** In Brassicaceae communities, polyploids tend to be more abundant than diploids, though the difference was not significant ($P > 0.05$). **(B)** In Rosaceae communities, diploids and polyploids do not significantly differ in abundance ($P > 0.05$). **(C)** In Brassicaceae communities, non-native (NN) polyploids tend to occur at a greater abundance than the other groups, but the difference is not significant ($P > 0.05$). **(D)** In Rosaceae communities, native species (N) are significantly more abundant than non-native species for both diploid and polyploid species ($P < 0.05$). Letters above the distributions in **(D)** indicate significantly different groups. The diamond and error bars indicate the mean of the distribution ± 1 standard error.

DISCUSSION

Polyploidy is now widely accepted as a mechanism of reproductive isolation and plant speciation, but much remains to be clarified about the influence of genome duplication on population- and community-level dynamics. In this study, we draw upon the extensive body of work conducted on the ecology and evolution of polyploids to predict and test how genome duplication may affect phylogenetic community structure. By examining two large flowering plant families with high incidences of polyploidy using phylogenetic data and cytogeographic information from a diversity of sources, we found that communities may be shaped in diverse ways by genome duplication and that the impacts of polyploidy are far from clear-cut. Polyploidy appears to influence patterns of phylogenetic relationships and species co-occurrence in Brassicaceae and Rosaceae communities, but these patterns appear to be lineage-specific rather than due to properties intrinsic to all genome duplication events. These results reflect the complexities and multifaceted consequences of polyploidy (Soltis et al., 2016), but our study also highlights the current paucity of information on ploidal variation at fine spatial scales (especially at cytotype contact zones), which may have contributed, in part, to some inconsistencies in our results.

Patterns of Polyploid Community Structure Are Lineage-Specific

For both Brassicaceae and Rosaceae, we found that ploidal variation is a common feature of communities across the United States. We especially observed a higher diversity of ploidal levels, and higher overall ploidies, among the Rosaceae. Of the 11 Rosaceae communities, all but one comprised both diploid and polyploid species, while two of the six Brassicacae communities were either composed of only polyploid species or of only diploid species. It is not immediately clear why Rosaceae species would exhibit a greater diversity of ploidies and higher ploidal complements, or why Rosaceae communities almost always included polyploids. This pattern may simply be due to the greater number of Rosaceae species present in the included communities, or that Rosaceae is an older family than Brassicaceae (∼95 vs. ∼42 million years old, respectively; Huang et al., 2016; Zhang et al., 2017) allowing more time for the evolution of greater ploidal diversity. However, it is notable that, compared to Brassicaceae, Rosaceae species tend to have perennial life histories. The longer-lived life histories of perennial species may satisfy conditions that promote unreduced gamete and polyploid formation, or polyploid phenotypes may best experience higher fitnesses when they have longer-lived life histories. Previous studies suggest that polyploid populations may arise more regularly in herbaceous species, but not necessarily in short-lived or annual species (Stebbins, 1938; Grant, 1981; Ramsey and Schemske, 2002; Zenil-Ferguson et al., 2017). It is possible that, on average, the Rosaceae species included in our analyses fall into a "sweet spot" of non-woody perennial life-history traits favoring genome duplication.

Our phylogenetic analyses of Brassicaceae and Rosaceae community structure indicate that in both families, polyploid species tend to be more distantly related to co-occurring diploids than diploids are to each other. Indeed, the proportion of diploid-diploid relationships falling below the MNTD of the family-level phylogeny was greater than that for the proportion of polyploid-diploid relationships (**Figure 3**). This suggests that the polyploid members of these communities may not have arisen *in situ*, but rather these polyploids are likely to have arisen in disjunct communities, or from interspecific hybridizations (i.e., allopolyploidy; Symonds et al., 2010), before dispersing to the surveyed communities. This is consistent with polyploids escaping minority cytotype exclusion by inhabiting differing geographic or ecological areas compared to their close relatives (Levin, 1975; Husband, 2000; Čertner et al., 2017). Alternatively, this phylogenetic pattern could have arisen if polyploids did establish within the same community as their diploid ancestors, but interploidal competition resulted in the local extinction of the diploid. Further studies incorporating a temporal aspect to community structure to capture interspecific interactions through time would allow us to distinguish between these two alternatives.

When considering each site separately, we did not find a consistent pattern in our simulation analyses. Although one Brassicaceae community showed polyploids to be more distantly related to diploids than expected by chance, at all other sites, we did not find a significant pattern, or found that polyploids were more closely related to diploids than expected, despite phylogenetic distances between polyploids and diploids being larger than the distances between diploids to one another. Together, the results from our broad-scale analyses and site-specific simulation analyses suggest that polyploidy can play an important role in shaping community structure but that the effect is species-specific. For example, the extent to which polyploids differ in phenotype and genetic composition could influence interactions with co-occurring species and the mode of escape from minority cytotype exclusion. Polyploids can exhibit wider variation in phenotypes compared to diploids, ranging from striking to subtle, which may depend in part upon the mode of polyploid formation. While allopolyploids often exhibit phenotypes that are intermediate to the parental species, the combination of two evolutionarily differentiated genomes, and their attendant regulatory elements, can sometimes produce transgressive phenotypes outside the range of variation harbored by either parental species (McCarthy et al., 2015, 2017). In contrast, autopolyploids often exhibit more subtle phenotypic differences when compared to their diploid progenitors (Maherali et al., 2009; Thompson et al., 2015). Therefore, our lack of a consistent result could be due, at least in part, to the inherent genetic differences between autopolyploids and allopolyploids, and further investigations examining how these two modes of polyploid formation may differ in their influence on community structure would go far toward illuminating interspecific interactions involving polyploids.

The apparent lineage-specific effect of polyploidy on phylogenetic community structure may also be due to varying ecological niche affinities and/or differences in life history. The hypothesized association between greater ploidal diversity and perennial life history (Müntzing, 1936; Stebbins, 1938; Grant, 1981) may mean that genome duplication shapes communities dominated by perennial species more strongly than communities comprising mostly annual species (Stebbins, 1938; Leitch and Leitch, 2012; Zenil-Ferguson et al., 2017). Though the incidence of polyploidy among woody species (which also tend to be perennial) is lower than among herbaceous species, this may consequently mean that the ecoregions or habitats dominated by perennial species (e.g., forests, woodlands, shrublands) are influenced more strongly by genome duplication than habitats where annual and herbaceous species dominate (e.g., grasslands, meadows). Our findings clearly provide motivation for broader investigations of the differences in impact on community structure between polyploid plant species with differing life histories.

Polyploids Not Consistently More Ecologically Successful Than Diploids

Polyploidy has classically been argued to be an important enabler of plant invasions and the exploitation of novel ecological niches (Pandit et al., 2014; Ramsey and Ramsey, 2014). Indeed, chromosome number has been identified as a correlate of invasiveness (Pyšek et al., 2013), and non-native polyploids in some flora are more likely to successfully become naturalized than diploid species (Nagy et al., 2017). As a measure of ecological success, a greater relative abundance of non-native species within a community should reflect their ability to successfully occupy and exploit novel habitat or outcompete and displace resident species (Levin, 1975; Callaway and Aschehoug, 2000; Cleland et al., 2004; te Beest et al., 2011). In our study, we found opposing patterns within Brassicaceae and Rosaceae communities. Brassicaceae polyploids showed patterns of abundance consistent with being more ecologically successful than diploids, which may have been driven by non-native species. Although there was not a significant difference between diploid and polyploid abundance, perhaps due to the relatively small sample of Brassicaceae sites analyzed, it is striking that in all communities that had both diploid and polyploid species, polyploids made up over 70% of the total relative abundance. On the other hand, Rosaceae diploid species were just as abundant as polyploids, and native species appeared to be more ecologically successful, with higher abundances, than the non-native species regardless of ploidy, suggesting that ecological success is not always a correlate of non-native and/or polyploid species.

The lack of a clear pattern for greater non-native polyploid abundance relative to diploids in Brassicaceae and Rosaceae communities is consistent with the varying findings of prior studies on invasive polyploids. For example, although many polyploids are invasive (Thompson, 1991; Pandit et al., 2006), species with smaller genome sizes have also been found to occur at higher species abundance, especially among annual species (Herben et al., 2012), and are more likely to be invasive (Grotkopp et al., 2002; Pandit et al., 2006, 2014; Kubešová et al., 2010; Lavergne et al., 2010; Herben and Goldberg, 2014; Schmidt et al., 2017). These counterintuitive findings may also reflect species-specific effects where a polyploid's potential for successful establishment and population expansion within a community may be highly dependent upon species-specific attributes, life histories, source locations, or the local environment of the community. For example, in anthropogenically disturbed habitats, non-native or invasive species are often polyploid (Lumaret and Borrill, 1988; Ramsey and Schemske, 1998). The importance of source locations and the ecology of the non-native range can also be seen in English Ivy (*Hedera* spp.), where the observation that diploids are invasive on the east coast of North America and tetraploids are invasive on the west coast of North America is thought to be due to adaptation that has occurred within the native European range, followed by subsequent exploitation of similar habitat within the invasive range (Green et al., 2013). Moreover, different cytotypes can also vary in ecological attributes and fitness across their range (McIntyre and Strauss, 2017), further nuancing the probability of establishment success within a community.

Conflicting observations of polyploid ecological success relative to diploids may also be due to the eco-evolutionary dynamics that occur over ecological timescales that affect interspecific competition and adaptation (Yoshida et al., 2003; Hairston et al., 2005; Reznick, 2013; DeLong et al., 2016). It is possible that when considered over time, the polyploid species observed in Brassicaceae and Roseaceae communities may be superior competitors that are in the process of displacing resident diploid species (or other ploidies). Alternatively, the polyploid species may be transient or ephemeral community members, documented at the present moment in time, and will eventually be displaced by the resident diploid species (Čertner et al., 2017). Additional studies incorporating phenotypic traits, and temporal data on species occurrence and abundance are needed to parse these alternatives and identify the underlying drivers of community structure. NEON's mission to repeatedly survey these sites over the next 30 years may provide an avenue to examine how community structure changes temporally, and offer insight into how polyploids and diploids interact within communities.

Observations that polyploids are not always ecologically superior suggest that polyploidy *per se* may have limited influence on the successful establishment of a population, or that the effects of genome duplication may not be uniformly predictable after polyploidy "primes the pump." This can be seen in studies explicitly examining ecological differences between diploids and polyploids that show variable patterns of ecological niche divergence for both auto- and allopolyploids (Glennon et al., 2014; Marchant et al., 2016). Studies involving synthetically generated polyploids have further demonstrated that interploidal trait differences only partially arise as a direct consequence of polyploidy, and similar studies in established polyploids are consistent with genome duplication either representing or generating intra-population variation that can be elaborated upon by natural selection (e.g., Husband and Schemske, 2000; Raabová et al., 2008; Ramsey, 2011; Laport et al., 2016). Additional studies incorporating ecological data (i.e., climate, soil, water availability, pollinators, etc.) would likely provide greater detail about diploid and polyploid differences at the community level in both native and non-native systems, and should be undertaken for a broader range of species (Kolár et al., 2017). Yet, additional comparative studies examining multiple diploid-polyploid pairs would go far in disentangling the influence of lineage- or cytotype-specific life history attributes, functional traits, and genomic contributions on the adaptive potential of genome duplication for range expansion and the establishment of non-native species within communities.

The Need for Greater Documentation of Geographic Ploidal Variation

Our study highlights the need for better documentation of intraspecific ploidal variation in a geographical context to better understand the role of genome duplication on plant community structure. Our characterization of members within a community was reliant upon local-scale documentation of ploidal variation, but we often found a paucity of available geographically explicit intraspecific ploidy data. Despite the known prevalence of geographic variation in ploidy within species (e.g., Baack, 2005; Kolár et al., 2009; Ståhlberg, 2009; Trávníček et al., 2011; Castro et al., 2012; Laport et al., 2012; Zozomová-Lihová et al., 2015; Wefferling et al., 2017; reviewed in Ramsey and Ramsey, 2014), species harboring populations differing in ploidy have historically been geographically under-sampled. Modern technologies, such as high throughput flow cytometry screening for DNA content (Kron et al., 2007), have improved our ability to identify intraspecific ploidal variation, representing potential cryptic biodiversity, and can facilitate tying phenotypic variation to different ploidies within polyploid complexes. Furthermore, new genomic tools and the ever-increasing trove of genomic data for non-model organisms could be used in *post-hoc* analyses to further reveal novel cytotype variation (e.g., modifications to genotype-by-sequencing approaches; Gompert and Mock, 2017). The implementation of these approaches, paired with broader usage of electronic databases (e.g., Kew C-value database, Chromosome Count Database) and inclusion of ploidy or genome size information on herbarium specimens will facilitate the documentation of polyploid complexes and further aid explorations of polyploid biodiversity and its influence on community structure.

CONCLUSION AND PERSPECTIVE

This is an exciting time to study the ecological and evolutionary implications of polyploidy at the population and community level. The growing body of work on polyploid evolution and population-level dynamics suggests that polyploidy may potentially have cascading effects on communities, yet few studies have explicitly tested the effect genome duplication has on community structure. Our novel study on Brassicaceae and Rosaceae communities suggests that the effects of genome duplication on community structure may often be lineage-specific, but polyploidy should still be considered as a potentially important driver of biodiversity patterns given the pervasiveness of genome duplication among vascular plants. Our findings contribute to the increasing number of studies highlighting the complexity and multifaceted consequences of whole genome duplication (reviewed in Ramsey and Ramsey, 2014; Soltis et al., 2016). Although explicitly population-level studies may reveal the processes underlying the pattern (e.g., inter-trophic-level interactions such as with herbivores, pollinators, mycorrhiza, and other microbial symbionts; reviewed in Segraves, 2017), macro-scale studies such as ours complement the many population-level studies of polyploids by providing a "zoomed out" perspective on general patterns, a comparative evaluation of a greater diversity of plant species and life histories, and offer nuance into how different evolutionary lineages may interact within communities comprising multiple ploidies.

At the same time, our understanding of the effect of polyploidy on community structure may have been hindered by the paucity of available geographically meaningful data on intraspecific ploidal variation, and the difficulty in compiling existing data from scattered literature reports. Alongside the recognized need to characterize intraspecific genetic and trait variation to understand their subsequent effects on community structure (Hughes et al., 2008; Bolnick et al., 2011), we urge continued emphasis on the characterization and documentation of ploidal variation across species' ranges. Such information will greatly aid comparative studies at the population and community level, and help shed light on how such a common, but profound, mutation influences the evolution of species and those around them.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the research and writing the paper. MG collected the data, and MG and JN performed statistical analyses.

FUNDING

This research was supported by a National Science Foundation grant to JN and RL (NSF-EF 1550813), an REU supplement to that grant, and a National Science Foundation grant to RL (NSF-DEB-1556371).

ACKNOWLEDGMENTS

We thank Vivianna Sanchez and William Weaver for their assistance in data collection and interpretation, and Chelsea Pretz for fruitful discussions in planning this study. We also thank Stacey D. Smith for generously sharing laboratory space to conduct this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2018.00052/full#supplementary-material

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001

Arvanitis, L., Wiklund, C., and Ehrlén, J. (2007). Butterfly seed predation: effects of landscape characteristics, plant ploidy level and population structure. *Oecologia* 152, 275–285. doi: 10.1007/s00442-007-0659-5

Baack, E. J. (2005). To succeed globally, disperse locally: effects of local pollen and seed dispersal on tetraploid establishment. *Heredity* 94, 538–546. doi: 10.1038/sj.hdy.6800656

Balao, F., Herrera, J., and Talavera, S. (2011). Phenotypic consequences of polyploidy and genome size at the microevolutionary scale: a multivariate morphological approach. *New Phytol.* 192, 256–265. doi: 10.1111/j.1469-8137.2011.03787.x

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A., and Knight, C. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.* 179, 975–986. doi: 10.1111/j.1469-8137.2008.02528.x

te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubešová, M., et al. (2011). The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* 109, 19–45. doi: 10.1093/aob/mcr277

Bilinski, P., Albert, P. S., Berg, J. J., Birchler, J., Grote, M., Lorant, A., et al. (2018). Parallel altitudinal clines reveal adaptive evolution of genome size in Zea mays. *bioRxiv*. doi: 10.1101/134528

Bolnick, D. I., Amarasekare, P., Araújo, M. S., Bürger, R., Levine, J. M., Novak, M., et al. (2011). Why intraspecific trait variation matters in community ecology. *Trends Ecol. Evol.* 26, 183–192. doi: 10.1016/j.tree.2011.01.009

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Callaway, R. M., and Aschehoug, E. T. (2000). Invasive plants versus their new and old neighbors: a mechanism for exotic invasion. *Science* 290, 521–523. doi: 10.1126/science.290.5491.521

Castro, S., Loureiro, J., Procházka, T., and Münzbergová, Z. (2012). Cytotype distribution at a diploid–hexaploid contact zone in *Aster amellus* (Asteraceae). *Ann. Bot.* 110, 1047–1055. doi: 10.1093/aob/mcs177

Cavender-Bares, J., Keen, A., and Miles, B. (2006). Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology* 87, S109–S122. doi: 10.1890/0012-9658(2006)87[109:PSOFPC]2.0.CO;2

Cleland, E. E., Smith, M. D., Andelman, S. J., Bowles, C., Carney, K. M., Horner-Devine, C. M., et al. (2004). Invasion in space and time: non-native species richness and relative abundance respond to interannual variation in productivity and diversity. *Ecol. Lett.* 7, 947–957. doi: 10.1111/j.1461-0248.2004.00655.x

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Coyne, J. A., and Orr, H. A. (2004). *Speciation.* Massachusetts, MA: Sinauer Associates, Inc.

Čertner, M., Fenclová, E., Kúr, P., Kolár, F., Koutecký, P., Krahulcová, A., et al. (2017). Evolutionary dynamics of mixed-ploidy populations in an annual herb: dispersal, local persistence and recurrent origins of polyploids. *Ann. Bot.* 120, 303–315. doi: 10.1093/aob/mcx032

DeLong, J. P., Forbes, V. E., Galic, N., Gibert, J. P., Laport, R. G., Phillips, J. S., et al. (2016). How fast is fast? Eco-evolutionary dynamics and rates of change in populations and phenotypes. *Ecol. Evol.* 6, 573–581. doi: 10.1002/ece3.1899

Doyle, J. D., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Emerson, B. C., and Gillespie, R. G. (2008). Phylogenetic analysis of community assembly and structure over space and time. *Trends Ecol. Evol.* 23, 619–630. doi: 10.1016/j.tree.2008.07.005

Fowler, N. L., and Levin, D. A. (2016). Critical factors in the establishment of allopolyploids. *Am. J. Bot.* 103, 1139–1145. doi: 10.3732/ajb.1500407

Glennon, K. L., Ritchie, M. E., and Segraves, K. A. (2014). Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecol. Lett.* 17, 574–582. doi: 10.1111/ele.12259

Gompert, Z., and Mock, K. E. (2017). Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Mol. Ecol. Resour.* 17, 1156–1167. doi: 10.1111/1755-0998.12657

Grant, V. (1981). *Plant Speciation.* New York, NY: Columbia University Press.

Green, A. F., Ramsey, T. S., and Ramsey, J. (2013). Polyploidy and invasion of English ivy (*Hedera spp*., Araliaceae) in North American forests. *Biol. Invasions* 15, 2219–2241. doi: 10.1007/s10530-013-0446-7

Grotkopp, E., Rejmánek, M., and Rost, T. L. (2002). Toward a causal explanation of plant invasiveness: seedling growth and life-history strategies of 29 pine (*Pinus*) species. *Am. Nat.* 159, 396–419. doi: 10.1086/338995

Hagberg, A., and Ellerström, S. (1959). The competition between diploid, tetraploid and aneuploid rye. *Hereditas* 45, 369–416. doi: 10.1111/j.1601-5223.1959.tb03058.x

Hairston, N. G., Ellner, S. P., Geber, M. A., Yoshida, T., and Fox, J. A. (2005). Rapid evolution and the convergence of ecological and evolutionary time. *Ecol. Lett.* 8, 1114–1127. doi: 10.1111/j.1461-0248.2005.00812.x

Halverson, K., Heard, S. B., Nason, J. D., and Stireman, J. O. (2008). Differential attack on diploid, tetraploid, and hexaploid *Solidago altissima* L. by five insect gallmakers. *Oecologia* 154, 755–761. doi: 10.1007/s00442-007-0863-3

Herben, T., and Goldberg, D. E. (2014). Community assembly by limiting similarity vs. competitive hierarchies: testing the consequences of dispersion of individual traits. *J. Ecol.* 102, 156–166. doi: 10.1111/1365-2745.12181

Herben, T., Suda, J., Klimešová, J., Mihulka, S., Ríha, P., and Šímová, I. (2012). Ecological effects of cell-level processes: genome size, functional traits and regional abundance of *herbaceous* plant species. *Ann. Bot.* 110, 1357–1367. doi: 10.1093/aob/mcs099

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., et al. (2016). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* 33, 394–412. doi: 10.1093/molbev/msv226

Hughes, A. R., Inouye, B. D., Johnson, M. T. J., Underwood, N., and Vellend, M. (2008). Ecological consequences of genetic diversity. *Ecol. Lett.* 11, 609–623. doi: 10.1111/j.1461-0248.2008.01179.x

Husband, B. C. (2000). Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. *Proc. R. Soc. Lond. B Biol. Sci.* 267, 217–223. doi: 10.1098/rspb.2000.0990

Husband, B. C., Ozimec, B., Martin, S. L., and Pollock, L. (2007). Mating consequences of polyploid evolution in flowering plants: current trends and insights from synthetic polyploids. *Int. J. Plant Sci.* 169, 195–206. doi: 10.1086/523367

Husband, B. C., and Schemske, D. W. (2000). Ecological mechanisms of reproductive isolation between diploid and tetraploid *Chamerion angustifolium. J. Ecol.* 88, 689–701. doi: 10.1046/j.1365-2745.2000.00481.x

Katoh, K., Misawa, K., Kuma, K. i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kay, Q. O. N. (1969). The origin and distribution of diploid and tetraploid *Tripleurospermum* i*nodorum* (L.) Schultz Bip. *Watsonia* 7, 130–141.

Keller, M., Schimel, D. S., Hargrove, W. W., and Hoffman, F. M. (2008). A continental strategy for the national ecological observatory network. *Front. Ecol. Environ.* 6, 282–284. doi: 10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., et al. (2010). Picante: r tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464. doi: 10.1093/bioinformatics/btq166

Kennedy, B. F., Sabara, H. A., Haydon, D., and Husband, B. C. (2006). Pollinator-mediated assortative mating in mixed ploidy populations of *Chamerion angustifolium* (Onagraceae). *Oecologia* 150, 398–408. doi: 10.1007/s00442-006-0536-7

Kolár, F., Certner, M., Suda, J., Schönswetter, P., and Husband, B.C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Plant Sci.* 22, 1041–1055. doi: 10.1016/j.tplants.2017.09.011

Kolár, F., Štech, M., Trávníček, P., Rauchová, J., Urfus, T., Vít, P., et al. (2009). Towards resolving the Knautia arvensis agg. *(Dipsacaceae) puzzle:* primary and secondary contact zones and ploidy segregation at landscape and microgeographic scales. *Ann. Bot.* 103, 963–974. doi: 10.1093/aob/mcp016

Koressaar, T., and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23, 1289–1291. doi: 10.1093/bioinformatics/btm091

Kreiner, J. M., Kron, P., and Husband, B. C. (2017). Frequency and maintenance of unreduced gametes in natural plant populations: associations with reproductive mode, life history and genome size. *New Phytol.* 214, 879–889. doi: 10.1111/nph.14423

Kron, P., Suda, J., and Husband, B. C. (2007). Applications of flow cytometry to evolutionary and population biology. *Ann. Rev. Ecol. Evol. Syst.* 38, 847–876. doi: 10.1146/annurev.ecolsys.38.091206.095504

Kubešová, M., Moravcova, L., Suda, J., Jarošík, V., and Pyšek, P. (2010). Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia* 82, 81–96.

Laport, R. G., Hatem, L., Minckley, R. L., and Ramsey, J. (2013). Ecological niche modeling implicates climatic adaptation, competitive exclusion, and niche conservatism among *Larrea tridentata* cytotypes in North American deserts. *J. Torrey Bot. Soc.* 140, 349–363. doi: 10.3159/TORREY-D-13-00009.1

Laport, R. G., Minckley, R. L., and Ramsey, J. (2012). Phylogeny and cytogeography of the North American creosote bush (*Larrea tridentata,* Zygophyllaceae). *Syst. Bot.* 37, 153–164. doi: 10.1600/036364412X616738

Laport, R. G., Minckley, R. L., and Ramsey, J. (2016). Ecological distributions, phenological isolation, and genetic structure in sympatric and parapatric populations of the *Larrea tridentata* polyploid complex. *Am. J. Bot.* 103, 1358–1374. doi: 10.3732/ajb.1600105

Laport, R. G., and Ng, J. (2017). Out of one, many: the biodiversity considerations of polyploidy. *Am. J. Bot.* 104, 1119–1121. doi: 10.3732/ajb.1700190

Lavergne, S., Mouquet, N., Thuiller, W., and Ronce, O. (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annu. Rev. Ecol. Evol. Syst.* 41, 321–350. doi: 10.1146/annurev-ecolsys-102209-144628

Leitch, A. R., and Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–483. doi: 10.1126/science.1153585

Leitch, A. R., and Leitch, I. J. (2012). Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* 194, 629–646. doi: 10.1111/j.1469-8137.2012.04105.x

Levin, D. A. (1975). Minority cytotype exclusion in local plant populations. *Taxon* 24, 35–43. doi: 10.2307/1218997

Levin, D. A. (1983). Polyploidy and novelty in flowering plants. *Am. Nat.* 122, 1–25. doi: 10.1086/284115

Levin, D. A., and Soltis, D. E. (2017). Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K-Pg interlude. *Curr. Opin. Plant Biol.* 42, 1–7. doi: 10.1016/j.pbi.2017.09.010

Lewis, W. H. (1962). Phylogenetic study of *Hedyotis* (Rubiaceae) in North America. *Am. J. Bot.* 49, 855–865. doi: 10.1002/j.1537-2197.1962.tb15020.x

Lumaret, R., and Borrill, M. (1988). Cytology, genetics, and evolution in the genus *dactylis*. *Crit. Rev. Plant Sci.* 7, 55–91. doi: 10.1080/07352688809382259

Maddison, W. P., and Maddison, D. R. (2017). *Mesquite: a modular system for evolutionary analysis*.Version 3.31. Available online at: http://mesquiteproject.org

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* 110, 99–104. doi: 10.1038/hdy.2012.79

Maherali, H., Walden, A. E., and Husband, B. C. (2009). Genome duplication and the evolution of physiological responses to water stress. *New Phytologist* 184, 721–731. doi: 10.1111/j.1469-8137.2009.02997.x

Manzaneda, A. J., Rey, P. J., Bastida, J. M., Weiss-Lehman, C., Raskin, E., and Mitchell-Olds, T. (2012). Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytol.* 193, 797–805. doi: 10.1111/j.1469-8137.2011.03988.x

Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytol.* 212, 708–718. doi: 10.1111/nph.14069

Masterson, J. (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264, 421–424. doi: 10.1126/science.264.5157.421

McCarthy, E. W., Arnold, S. E., Chittka, L., Le Comber, S. C., Verity, R., Dodsworth, S., et al. (2015). The effect of polyploidy and hybridization on the evolution of floral colour in *Nicotiana* (Solanaceae). *Ann. Bot.* 115, 1117–1131. doi: 10.1093/aob/mcv048

McCarthy, E. W., Berardi, A. E., Smith, S. D., and Litt, A. (2017). Related allopolyploids display distinct floral pigment profiles and transgressive pigments. *Am. J. Bot.* 104, 92–101. doi: 10.3732/ajb.1600350

McIntyre, P. J. (2012). Polyploidy associated with altered and broader ecological niches in the *Claytonia perfoliata* (Portulacaceae) species complex. *Am. J. Bot.* 99, 655–662. doi: 10.3732/ajb.1100466

McIntyre, P. J., and Strauss, S. (2017). An experimental test of local adaptation among cytotypes within a polyploid complex. *Evolution* 71, 1960–1969. doi: 10.1111/evo.13288

Mei, W., Stetter, M. G., Gates, D. J., Stitzer, M., and Ross-Ibarra, J. (2018). Adaptation in plant genomes: bigger is different. *Am. J. Bot.* 105, 16–19. doi: 10.1002/ajb2.1002

Müntzing, A. (1936). The evolutionary significance of autopolyploidy. *Hereditas* 21, 363–378. doi: 10.1111/j.1601-5223.1936.tb03204.x

Nagy, D. U., Stranczinger, S., Godi, A., Weisz, A., Rosche, C., Suda, J., et al. (2017). Does higher ploidy level increase the risk of invasion? A case study with two geo-cytotypes of *Solidago gigantea* Aiton (Asteraceae). *J. Plant Ecol.* 11, 317–327. doi: 10.1093/jpe/rtx005

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462. doi: 10.1016/j.cell.2007.10.022

Pandit, M. K., Tan, H. T. W., and Bisht, M. S. (2006). Polyploidy in invasive plant species of Singapore. *Bot. J. Linnean Soc.* 151, 395–403. doi: 10.1111/j.1095-8339.2006.00515.x

Pandit, M. K., White, S. M., and Pocock, M. J. O. (2014). The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. *New Phytol.* 203, 697–703. doi: 10.1111/nph.12799

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Park, D. S., Worthington, S., and Xi, Z. (2018). Taxon sampling effects on the quantification and comparison of community phylogenetic diversity. *Mol. Ecol.* 27, 1296–1308. doi: 10.1111/mec.14520

Pyšek, P., Hulme, P. E., Meyerson, L. A., Smith, G. F., Boatwright, J. S., Crouch, N. R., et al. (2013). Hitting the right target: taxonomic challenges for, and of, plant invasions. *AoB Plants* 5:plt042. doi: 10.1093/aobpla/plt042

Raabová, J., Fischer, M., and Münzbergová, Z. (2008). Niche differentiation between diploid and hexaploid *Aster amellus*. *Oecologia* 158, 463–472. doi: 10.1007/s00442-008-1156-1

Rambaut, A., Suchard, M., Xie, D., and Drummond, A. (2014). *Tracer* v1. 6 Available online at: http://beast.bio.ed.ac.uk

Ramsey, J. (2011). Polyploidy and ecological adaptation in wild yarrow. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7096–7101. doi: 10.1073/pnas.1016631108

Ramsey, J., and Ramsey, T. S. (2014). Ecological studies of polyploidy in the 100 years following its discovery. *Philos. Trans. R. Soc. B* 369:20130352. doi: 10.1098/rstb.2013.0352

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Ramsey, J., and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* 33, 589–639. doi: 10.1146/annurev.ecolsys.33.010802.150437

Rey, P. J., Manzaneda, A. J., and Alcantara, J. M. (2017). The interplay between aridity and competition determines colonization ability, exclusion, and ecological segregation in the heteroploid *Brachypodium distachyon* species complex. *New Phytol.* 215, 85–96. doi: 10.1111/nph.14574

Reznick, D. N. (2013). A critical look at reciprocity in ecology and evolution: introduction to the symposium. *Am. Nat.* 181, S1–S8. doi: 10.1086/670030

Rieseberg, L. H., and Willis, J. H. (2007). Plant speciation. *Science* 317, 910–914. doi: 10.1126/science.1137729

Roccaforte, K., Russo, S. E., and Pilson, D. (2015). Hybridization and reproductive isolation between diploid *Erythronium mesochoreum* and its tetraploid congener *E. albidum (Liliaceae). Evolution* 69, 1375–1389. doi: 10.1111/evo.12666

Schlaepfer, D. R., Edwards, P. J., and Billeter, R. (2010). Why only tetraploid *Solidago gigantea* (Asteraceae) became invasive: a common garden comparison of ploidy levels. *Oecologia* 163, 661–673. doi: 10.1007/s00442-010-1595-3

Schmidt, J. P., Drake, J. M., and Stephens, P. (2017). Residence time, native range size, and genome size predict naturalization among angiosperms introduced to Australia. *Ecol. Evol.* 7, 10289–10300. doi: 10.1002/ece3.3505

Segraves, K. A. (2017). The effects of genome duplications in a community context. *New Phytol.* 215, 57–69. doi: 10.1111/nph.14564

Segraves, K., and Thompson, J. (1999). Plant polyploidy and pollination: floral traits and insect visits to diploid and tetraploid *Heuchera grossulariifolia*. *Evolution* 53, 1114–1127. doi: 10.1111/j.1558-5646.1999.tb04526.x

Smith, S. A., Beaulieu, J. M., and Donoghue, M. J. (2009). Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol. Biol.* 9:37. doi: 10.1186/1471-2148-9-37

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., et al. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. doi: 10.3732/ajb.0800079

Soltis, D. E., Soltis, P. S., Schemske, D. W., Hancock, J. F., Thompson, J. N., Husband, B. C., et al. (2007). Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56, 13–30. doi: 10.2307/25065732

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Soltis, P. S., Liu, X., Marchant, D. B., Visger, C. J., and Soltis, D. E. (2014). Polyploidy and novelty: Gottlieb's legacy. *Philos. Trans. R. Soc. B* 369:20130351. doi: 10.1098/rstb.2013.0351

Soltis, P. S., Marchant, D. B., Van de Peer, Y., and Soltis, D. E. (2015). Polyploid and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125. doi: 10.1016/j.gde.2015.11.003

Song, Q., and Chen, Z. J. (2015). Epigenetic and developmental regulation in plant polyploids. *Curr. Opin. Plant Biol.* 24, 101–109. doi: 10.1016/j.pbi.2015.02.007

Ståhlberg, D. (2009). Habitat differentiation, hybridization and gene flow patterns in mixed populations of diploid and autotetraploid *Dactylorhiza maculata* s.l. (Orchidaceae). *Evol. Ecol.* 23, 295–328. doi: 10.1007/s10682-007-9228-y

Stebbins, G. L. (1938). Cytological characteristics associated with the different growth habits in the Dicotyledons. *Am. J. Bot.* 25, 189–198. doi: 10.1002/j.1537-2197.1938.tb09203.x

Symonds, V. V., Soltis, P. S., and Soltis, D. E. (2010). Dynamics of polyploid formation in *Tragopogon* (Asteraceae): recurrent formation, gene flow, and population structure. *Evolution* 64, 1984–2003. doi: 10.1111/j.1558-5646.2010.00978.x

Těšitelová, T., Jersáková, J., Roy, M., Kubátová, B., Těšitel, J., Urfus, T., et al. (2013). Ploidy-specific symbiotic interactions: divergence of mycorrhizal fungi between cytotypes of the *Gymnadenia conopsea* group (Orchidaceae). *New Phytol.* 199, 1022–1033. doi: 10.1111/nph.12348

Thompson, J. D. (1991). The biology of an invasive plant. *BioScience* 41, 393–401. doi: 10.2307/1311746

Thompson, J. D., and Lumaret, R. (1992). The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends Ecol. Evol.* 7, 302–307. doi: 10.1016/0169-5347(92)90228-4

Thompson, J. N., Cunningham, B. M., Segraves, K. A., Althoff, D. M., and Wagner, D. (1997). Plant polyploidy and insect/plant interactions. *Am. Nat.* 150, 730–743. doi: 10.1086/286091

Thompson, J. N., and Merg, K. F. (2008). Evolution of polyploidy and the diversification of plant-pollinator interactions. *Ecology* 89, 2197–2206. doi: 10.1890/07-1432.1

Thompson, K. A., Husband, B. C., and Maherali, H. (2015). No influence of water limitation on the outcome of competition between diploid and tetraploid *Chamerion angustifolium* (Onagraceae). *J. Ecol.* 103, 733–741. doi: 10.1111/1365-2745.12384

Trávníček, P., Kubátová, B., Curn, V., Rauchová, J., Krajníková, E., Jersáková, J., et al. (2011). Remarkable coexistence of multiple cytotypes of the *Gymnadenia conopsea* aggregate (the fragrant orchid): evidence from flow cytometry. *Ann. Bot.* 107, 77–87. doi: 10.1093/aob/mcq217

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

Vamosi, S. M., Heard, B., Vamosi, J. C., and Webb, C. O. (2009). Emerging patterns in the comparative analysis of phylogenetic community structure. *Mol. Ecol.* 18, 572–592. doi: 10.1111/j.1365-294X.2008.04001.x

Webb, C. O., Ackerly, D. D., McPeek, M. A., and Donoghue, M. J. (2002). Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* 33, 475–505. doi: 10.1146/annurev.ecolsys.33.010802.150448

Wefferling, K. M., Castro, S., Loureiro, J., Castro, M., Tavares, D., and Hoot, S. B. (2017). Cytogeography of the subalpine marsh marigold polyploid complex (*Caltha leptosepala* s.l., Ranunculaceae). *Am. J. Bot.* 104, 271–285. doi: 10.3732/ajb.1600365

Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F., and Hairston, N. G. Jr. (2003). Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* 424, 303–306. doi: 10.1038/nature01767

Zenil-Ferguson, R., Ponciano, J. M., and Burleigh, J. G. (2017). Testing the association of phenotypes with polyploidy: an example using herbaceous and woody eudicots. *Evolution* 71, 1138–1148. doi: 10.1111/evo.13226

Zhan, S. H., Drori, M., Goldberg, E. E., Otto, S. P., and Mayrose, I. (2016). Phylogenetic evidence for cladogenetic polyploidization in land plants. *Am. J. Bot.* 103, 1252–1258. doi: 10.3732/ajb.1600108

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., et al. (2017). Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. *New Phytol.* 214, 1355–1367. doi: 10.1111/nph.14461

Zozomová-Lihová, J., Malánová-Krásná, I., Vít, P., Urfus, T., Senko, D., Svitok, M., et al. (2015). Cytotype distribution patterns, ecological differentiation, and genetic structure in a diploid–tetraploid contact zone of *Cardamine amara*. *Am. J. Bot.* 102, 1380–1395. doi: 10.3732/ajb.1500052

Conflict of Interest Statement: 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: 06 August 2018
doi: 10.3389/fgene.2018.00296

Spindle Dynamics Model Explains Chromosome Loss Rates in Yeast Polyploid Cells

Ivan Jelenić [1], Anna Selmecki [2], Liedewij Laan [3] and Nenad Pavin [1]**

[1] Department of Physics, Faculty of Science, University of Zagreb, Zagreb, Croatia, [2] Department of Medical Microbiology and Immunology, Creighton University Medical School, Omaha, NE, United States, [3] Department of Bionanoscience, Faculty of Applied Sciences, Kavli Institute of NanoScience, Delft University of Technology, Delft, Netherlands

OPEN ACCESS

Edited by:
Richard John Abbott,
University of St Andrews,
United Kingdom

Reviewed by:
Pirita Paajanen,
John Innes Centre (JIC),
United Kingdom
Yongshuai Sun,
Xishuangbanna Tropical Botanical
Garden (CAS), China

***Correspondence:**
Liedewij Laan
l.laan@tudelft.nl
Nenad Pavin
npavin@phy.hr

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

Received: 02 March 2018
Accepted: 13 July 2018
Published: 06 August 2018

Citation:
Jelenić I, Selmecki A, Laan L and
Pavin N (2018) Spindle Dynamics
Model Explains Chromosome Loss
Rates in Yeast Polyploid Cells.
Front. Genet. 9:296.
doi: 10.3389/fgene.2018.00296

Faithful chromosome segregation, driven by the mitotic spindle, is essential for organismal survival. Neopolyploid cells from diverse species exhibit a significant increase in mitotic errors relative to their diploid progenitors, resulting in chromosome nondisjunction. In the model system *Saccharomyces cerevisiae,* the rate of chromosome loss in haploid and diploid cells is measured to be one thousand times lower than the rate of loss in isogenic tetraploid cells. Currently it is unknown what constrains the number of chromosomes that can be segregated with high fidelity in an organism. Here we developed a simple mathematical model to study how different rates of chromosome loss in cells with different ploidy can arise from changes in (1) spindle dynamics and (2) a maximum duration of mitotic arrest, after which cells enter anaphase. We apply this model to *S. cerevisiae* to show that this model can explain the observed rates of chromosome loss in *S. cerevisiae* cells of different ploidy. Our model describes how small increases in spindle assembly time can result in dramatic differences in the rate of chromosomes loss between cells of increasing ploidy and predicts the maximum duration of mitotic arrest.

Keywords: polyploidy, spindle assembly, chromosome loss, chromosome segregation, cell cycle regulation, theoretical modeling, genome instability

INTRODUCTION

Chromosome segregation is an important, highly conserved cellular function. A complex network of interacting components segregates chromosomes with high precision. However, rare errors in chromosome segregation are observed, and the error rate generally increases when the number of sets of chromosomes (ploidy, *n*) increases within the cell (Comai, 2005). Increased rates of chromosome loss are observed in autopolyploid cells, within yeasts, plants, and human cells (Mayer and Aguilera, 1990; Song et al., 1995; Ganem et al., 2009). For example, autopolyploidization of *Phlox drummondii* results in an immediate loss of approximately 17% of genomic DNA in the first generation and up to 25% after three generations (Raina et al., 1994). Autopolyploidization can also cause tumorigenesis, and these tumors are marked by significant chromosome gain/loss events (Fujiwara et al., 2005; Zack et al., 2013). Therefore, the general observation is that many newly formed polyploid cells have increased chromosome segregation errors relative to isogenic diploid cells, and the cause of these errors is not known.

The normal sexual life cycle of the budding yeast *Saccharomyces cerevisiae* includes haploid ($n = 1$, 16 chromosomes) and diploid cells ($n = 2$, 32 chromosomes). In addition, tetraploid cells ($n = 4$, 64 chromosomes) are rarely found in nature, but can be generated in the lab by mating two diploid cells. In this organism, the effect of ploidy on the rate of chromosome loss is very pronounced: haploid and diploid cells have rates of chromosome loss around 10^{-6} chromosomes per cell per cell division, whereas tetraploid cells have a rate around 10^{-3} (Mayer and Aguilera, 1990; Storchová et al., 2006). The rate of chromosome loss was measured with isogenic haploid, diploid, and tetraploid strains that each contained a single genetically marked chromosome. In these assays the cells that have lost the chromosome markers are quantified, and the rate of loss is determined by fluctuation analysis (Lea and Coulson, 1949). Moreover, polyploid laboratory yeast strains tend to lose chromosomes and reduce to a diploid level in experimental evolution studies (Gerstein et al., 2006; Selmecki et al., 2015). Thus, the genomic stability of a cell line is to a large extent related to cellular ploidy, but how ploidy alters chromosome segregation is not known (Otto and Whitton, 2000).

Chromosome segregation is driven by the mitotic spindle, a self-organized micro-machine composed of microtubules and associated proteins (Pavin and Tolić, 2016; Prosser and Pelletier, 2017). In budding yeast, during spindle assembly, spindle poles nucleate microtubules, which grow in a direction parallel with the central spindle or in arbitrary directions within the nucleus (Winey et al., 1995; O'Toole et al., 1997). A microtubule that comes into the proximity of a kinetochore (KC), a protein complex at the sister chromatids, can attach to the KC and thus establish a link between chromatids and spindle poles, as shown *in vitro* (Mitchison and Kirschner, 1985; Akiyoshi et al., 2010; Gonen et al., 2012; Volkov et al., 2013), *in vivo* (Tanaka et al., 2005), and theoretically (Hill, 1985). Theoretical models have quantitatively shown that this process can contribute to spindle assembly in yeasts and in mammalian cells (Wollman et al., 2005; Paul et al., 2009; Kalinina et al., 2013; Vasileva et al., 2017). Prior to chromosome separation, all connections between chromatids and the spindle pole must be established, and erroneous KC-microtubule attachments must be corrected, for which several theoretical models have been proposed (Zaytsev and Grishchuk, 2015; Tubman et al., 2017). These connections are monitored by the spindle assembly checkpoint (Li and Murray, 1991). Once KCs are properly attached and chromosomes congress to the metaphase plate (Gardner et al., 2008), the spindle assembly checkpoint is silenced and microtubules separate the sister chromatids (Musacchio and Salmon, 2007).

Cells that cannot satisfy the spindle assembly checkpoint are arrested in mitosis. However, cells can break out of the arrest after several hours, an event that is often referred to as "mitotic slippage" (Minshull et al., 1996; Rudner and Murray, 1996; Rieder and Maiato, 2004), and this mitotic exit is molecularly regulated (Novák et al., 1999; Rudner et al., 2000). Even though the molecular mechanisms that regulate cell cycle and spindle assembly are emerging, it is an open question as to how changes in ploidy can have such a dramatic effect on the rates of chromosome loss.

In this paper, we introduce a theoretical model for chromosome loss in cells with different ploidy. We test the hypothesis that polyploidy limits faithful chromosome segregation by the combination of dynamics of spindle assembly and a maximum time of mitotic arrest. Our model predicts that for increasing ploidy, spindle assembly time scales linearly with the number of chromosomes, which results in exponential changes in the rate of chromosome loss. Our model quantitatively reproduces the increase in chromosome loss observed in tetraploid *S. cerevisiae* cells relative to haploid and diploid cells.

MATERIALS AND METHODS

Model for Chromosome Loss

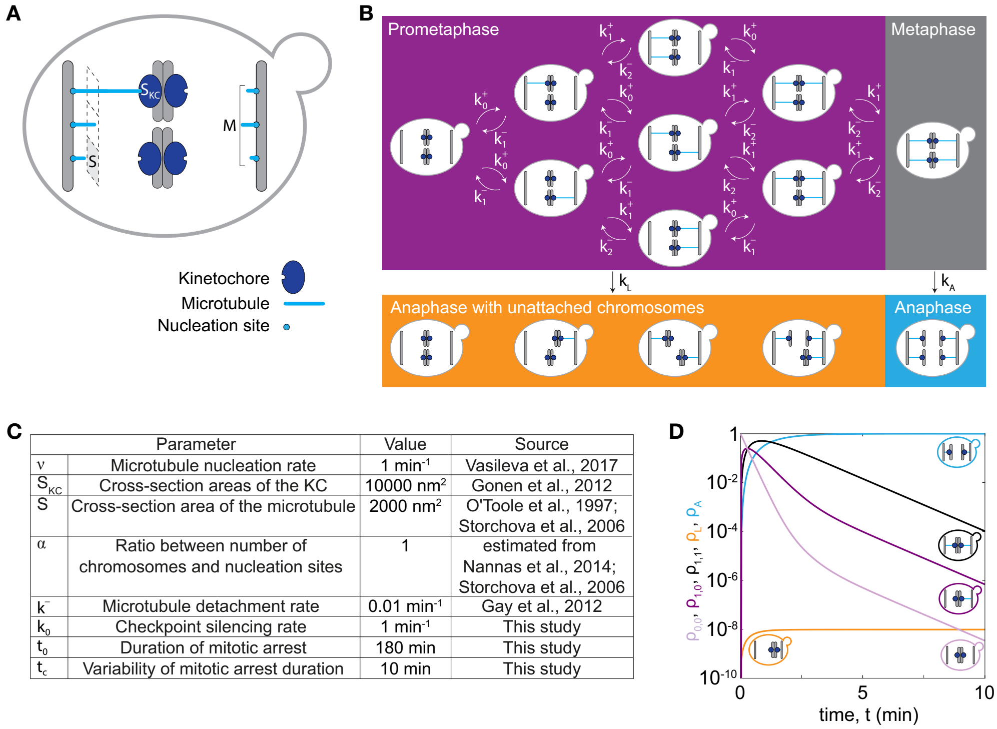
In our model we describe the dynamics of spindle assembly including KC attachment and detachment (**Figure 1A**), silencing of the spindle assembly checkpoint and the maximum duration of mitotic arrest after which cells enter anaphase regardless of whether all KCs are attached, allowing for chromosome loss in our model. To make a prediction for chromosome loss, we describe populations of cells in prometaphase, metaphase, and anaphase with either all KCs attached to the spindle, or with at least one unattached KC, and we calculate the fraction of cells in each population (**Figure 1B**). Transitions between these populations arise from spindle dynamics (**Figure 1A**).

Dynamics of Spindle Assembly

To describe dynamics of spindle assembly, we calculate the rate of KC capture, k_i^+, by taking into account known microtubule dynamics and geometry of yeast spindles (**Figure 1A**). Here, index i denotes the number of left sister KCs attached to the spindle; analogous calculations are applied to right sister KCs. Microtubules nucleate from the spindle pole body at rate ν_i and extend toward the spindle equator. They can attach to an unattached KC with probability p. The rate of KC attachment is the probability of attachment of one of the unattached KCs multiplied with the microtubule nucleation rate, which for C chromosomes and $C - i$ unattached KCs reads

$$k_i^+ = \left[1 - \left(1 - p\right)^{C-i}\right] \nu_i, \quad i = 0, \ldots, C - 1. \tag{1}$$

For other values of the index i the rate of KC attachment is zero to exclude unrealistic cases, with a negative number of chromosomes or with more than C chromosomes. In the case of euploid cells, the number of chromosomes is related to the ploidy as $C = 16 \cdot n$. We calculate the nucleation rate at the spindle pole body as $\nu_i = \nu \cdot (M{-}i)$, where we assume that a spindle pole body has a constant number of M nucleation sites with $M - i$ unoccupied nucleation sites. To determine M for different numbers of chromosomes, we introduce a linear relationship between the number of chromosomes and nucleation sites, $M = \alpha \cdot C + 4$, which is based on experimental findings (Storchová et al., 2006; Nannas et al., 2014). The parameter α is typically around 1. We also assume the nucleation rate for one nucleation site, ν , to be constant as in previous studies (Kitamura et al., 2010; Vasileva et al., 2017). In our model, attachment occurs when a microtubule contacts the KC (Tanaka et al., 2005). The



	Parameter	Value	Source
ν	Microtubule nucleation rate	1 min^{-1}	Vasileva et al., 2017
S_{KC}	Cross-section areas of the KC	10000 nm^2	Gonen et al., 2012
S	Cross-section area of the microtubule	2000 nm^2	O'Toole et al., 1997; Storchova et al., 2006
α	Ratio between number of chromosomes and nucleation sites	1	estimated from Nannas et al., 2014; Storchova et al., 2006
k^-	Microtubule detachment rate	0.01 min^{-1}	Gay et al., 2012
k_0	Checkpoint silencing rate	1 min^{-1}	This study
t_0	Duration of mitotic arrest	180 min	This study
t_c	Variability of mitotic arrest duration	10 min	This study

FIGURE 1 | Model for chromosome loss. **(A)** Spindle geometry in an individual cell. A microtubule (light blue) occupies a cross-section area *S*. Microtubules nucleate from *M* nucleation sites at the spindle pole body (gray bar) and extend toward KCs (dark blue) of a cross-section area S_{KC}. **(B)** Spindle dynamics in mitosis. The different boxes indicate cells in prometaphase (purple box), metaphase (gray box) and anaphase (orange and blue box). Arrows denote the rate of transition between different populations. Within a cell, microtubules (blue lines) extend from the spindle pole bodies (gray bars) toward the KCs (dark blue circles). **(C)** Parameters used to solve the model. Five parameter values were taken from previous studies (O'Toole et al., 1997; Storchová et al., 2006; Gay et al., 2012; Gonen et al., 2012; Nannas et al., 2014; Vasileva et al., 2017), as indicated. **(D)** Solution of the model for cells with 1 chromosome ($C = 1$). Fraction of cells in prometaphase with no KCs attached (light purple, $\rho_{0,0}$), with 1 KC attached (dark purple, $\rho_{1,0}$ or $\rho_{0,1}$), in metaphase (black, $\rho_{1,1}$), in anaphase with at least one KC unattached (orange, ρ_L) and in anaphase (blue, ρ_A), are shown. Each line is accompanied by a cell cartoon depicting the corresponding phase of the cell cycle. At $t = 0$, $\rho_{0,0} = 1$ and all other populations are 0.

probability of attachment is calculated based on spindle geometry as the ratio of the cross-section areas of the KC, S_{KC}, and the total area of the spindle, $p = S_{KC}/(S \cdot M + S_{KC})$. Here S denotes the cross-section area occupied by one microtubule. Values for these parameters are estimated from electron microscopy studies (O'Toole et al., 1997; Storchová et al., 2006; Gonen et al., 2012). We assume that microtubules detach from one KC at constant detachment rate, k^-, because our model does not include forces at the KC (Akiyoshi et al., 2010).

Silencing the Spindle Assembly Checkpoint and Chromosome Loss

Cells proceed from metaphase to anaphase by silencing the spindle assembly checkpoint at a constant rate, k_0. They can also proceed from prometaphase to anaphase when they spend a prolonged time in mitotic arrest (Minshull et al., 1996; Rudner and Murray, 1996; Rieder and Maiato, 2004), which in our model results in chromosome loss. We distinguish these two cases by introducing a rate of anaphase entry given by

$$\begin{Bmatrix} k_L \\ k_A \end{Bmatrix} = k_0 \begin{Bmatrix} f(t) \\ 1 + f(t) \end{Bmatrix}, \qquad (2)$$

where in the top and bottom row we calculate rates at which cells leave prometaphase and metaphase, respectively. We describe bypassing the checkpoint in mitotic arrest with a function of time $f(t)$, irrespective whether cells are in prometaphase or metaphase. Because this function is not known, we choose a simple mathematical form $f(t) = exp[(t - t_0)/t_c]$, which accounts for the rate of anaphase entry increase in time. Here, parameters t_0 and t_c denote the duration of mitotic arrest and the characteristic timescale, respectively.

Fraction of Cells in Prometaphase, Metaphase, and Anaphase With and Without Lost Chromosomes

In our model, we denote the fractions of cells in prometaphase and metaphase by $\rho_{i,j}$. The fraction of cells in anaphase with at least one KC unattached to the spindle, ρ_L, represents the fraction with lost chromosomes. The fraction of cells in anaphase with all KCs attached is denoted ρ_A. The indices i and j denote the number of left and right sister KCs attached to the spindle, respectively, in cells with C chromosomes ($i = 0, \ldots, C$ and $j = 0, \ldots, C$). The combination of indices $i = j = C$ describes cells with all KCs attached, which corresponds to metaphase cells. All the other combinations of indices describe cells with at least one unattached KC, which correspond to prometaphase cells. As time, t, progresses (i) KCs attach to or detach from the spindle, or (ii) cells enter anaphase changing the factions of cells in the populations (**Figure 1B**). In our model, attachments of different KCs as well as their detachments are independent. We describe these processes by a system of rate equations:

$$\begin{aligned}\frac{d\rho_{i,j}}{dt} &= k^{+}_{i-1}\rho_{i-1,j} + k^{+}_{j-1}\rho_{i,j-1} + (i+1)k^{-}\rho_{i+1,j} \\ &\quad +(j+1)k^{-}\rho_{i,j+1} - (k^{+}_{i} + ik^{-} + k^{+}_{j} + jk^{-} \\ &\quad +k_{\mathrm{L,A}})\rho_{i,j}, \quad i,j = 0,\ldots,C \qquad (3)\\ k_{\mathrm{L,A}} &= \begin{cases} k_{\mathrm{A}}, \text{if } i = j = C \\ k_{\mathrm{L}} \text{ otherwise} \end{cases}, \\ \frac{d\rho_{\mathrm{L}}}{dt} &= k_{\mathrm{L}} \sum_{i,j=0}^{C} \rho_{i,j}(1-\delta_{i,C}\delta_{j,C}), \qquad (4)\\ \frac{d\rho_{\mathrm{A}}}{dt} &= k_{\mathrm{A}}\rho_{C,C}. \qquad (5)\end{aligned}$$

Here δ denotes the Kronecker delta function, which has value 1 when two indices have the same value and 0 otherwise. Note that equation (3) describes a situation where only one KC can attach to or detach from the spindle at a time, which can be used if KCs attach and detach independently of each other. We also introduce the average time of both prometaphase and metaphase, which we term the time of spindle assembly, $\langle t\rangle = \int_{0}^{\infty} t\frac{d\rho_A}{dt}dt / \int_{0}^{\infty} \frac{d\rho_A}{dt}dt$. Please note that the model does not take cell division into account and therefore the total number of cells is conserved.

RESULTS

Chromosome Loss in Cells With One Chromosome

To illustrate how chromosome loss occurs during the transition from prometaphase to anaphase, we numerically solve our model first for cells with only one chromosome, $C = 1$, for parameters given in **Figure 1C**. We discuss the time course for different populations of cells. Initially, cells have no chromosome attached to the spindle. In prometaphase, when spindle assembly starts and KCs attach to the spindle, the fraction of cells in this population decreases, while the fraction of cells in the other populations increases (compare the light and dark purple lines in **Figure 1D**). After an initial increase, the fraction of cells in prometaphase starts decreasing as more KCs attach, and cells switch to metaphase (compare purple and black lines in **Figure 1D**). Finally, cells switch to anaphase. The fractions of cells in anaphase increase and asymptotically approach a limit value because the model does not describe cells leaving anaphase (orange and blue lines in **Figure 1D**). In this case with only one chromosome, the fraction of cells with a lost chromosome is very low.

Dramatic Increase in the Rate of Chromosome Loss With an Increase in Ploidy

To explore the relevance of our model for haploid, diploid, and tetraploid yeast cells, we further solve our model for the respective number of chromosomes in each ploidy type, $C = 16$, 32, and 64 (**Figure 2A**). We find that cells with an increasing number of chromosomes spend a longer time in prometaphase and metaphase, though the general trend is similar to the case with $C = 1$ (**Figure 1D**). Additionally, there is a rapid decrease in the fraction of cells in prometaphase and metaphase, which occurs around the maximum time of mitotic arrest, $t = t_0$, which is visible for cells with 64 chromosomes. After cells pass the maximum time of mitotic arrest, they predominantly enter anaphase regardless whether all KCs are attached. Thus, the more cells are still in prometaphase, the more cells will enter anaphase with unattached KCs. Because populations of cells with more chromosomes spend more time in prometaphase, they also enter anaphase later (**Figures 2A,B**). This time delay results in an increasing fraction of cells in anaphase with at least one lost KC because these cells have a greater chance to proceed to anaphase without a completely formed spindle (**Figure 2B**).

To explore which processes included in our model are responsible for significant chromosome loss, we determine the relevance of our model parameters. As our model describes both KC capture and transition to anaphase, we separately analyse the contribution of each process. We introduce the average time of both prometaphase and metaphase, which we refer to as the time of spindle assembly (Methods). We find that the time of spindle assembly increases with the number of chromosomes. Changing the chromosome number from 16 to 32 increases the time of spindle assembly approximately 2-fold, whereas, for a change from 32 to 64, it increases 5-fold (**Figure 2C**). Next, we explored how ploidy variations affect chromosome loss. We find that haploid ($C = 16$) and diploid ($C = 32$) cells have the same order of magnitude for the fraction of the population with at least one lost chromosome (**Figure 2D**). Interestingly, the fraction of cells with at least one lost chromosome increases dramatically for cells with higher ploidy, such as tetraploid cells ($C = 64$). When we plot the fraction of cells with lost kinetochores against spindle assembly time, we find that linear-scale changes in spindle assembly time result in exponential-scale changes in the rate of chromosome loss (**Figure 2E**). To summarize, our combined results show that small changes in spindle assembly time result in dramatic differences in the rate of chromosome loss as soon as prometaphase time approaches the maximum time of mitotic arrest.

Relevance of Parameters on the Time of Spindle Assembly and the Chromosome Loss Rate

As our model describes spindle formation, we explore the relevance of parameters on the time of spindle assembly. We varied the parameter that links the number of chromosomes

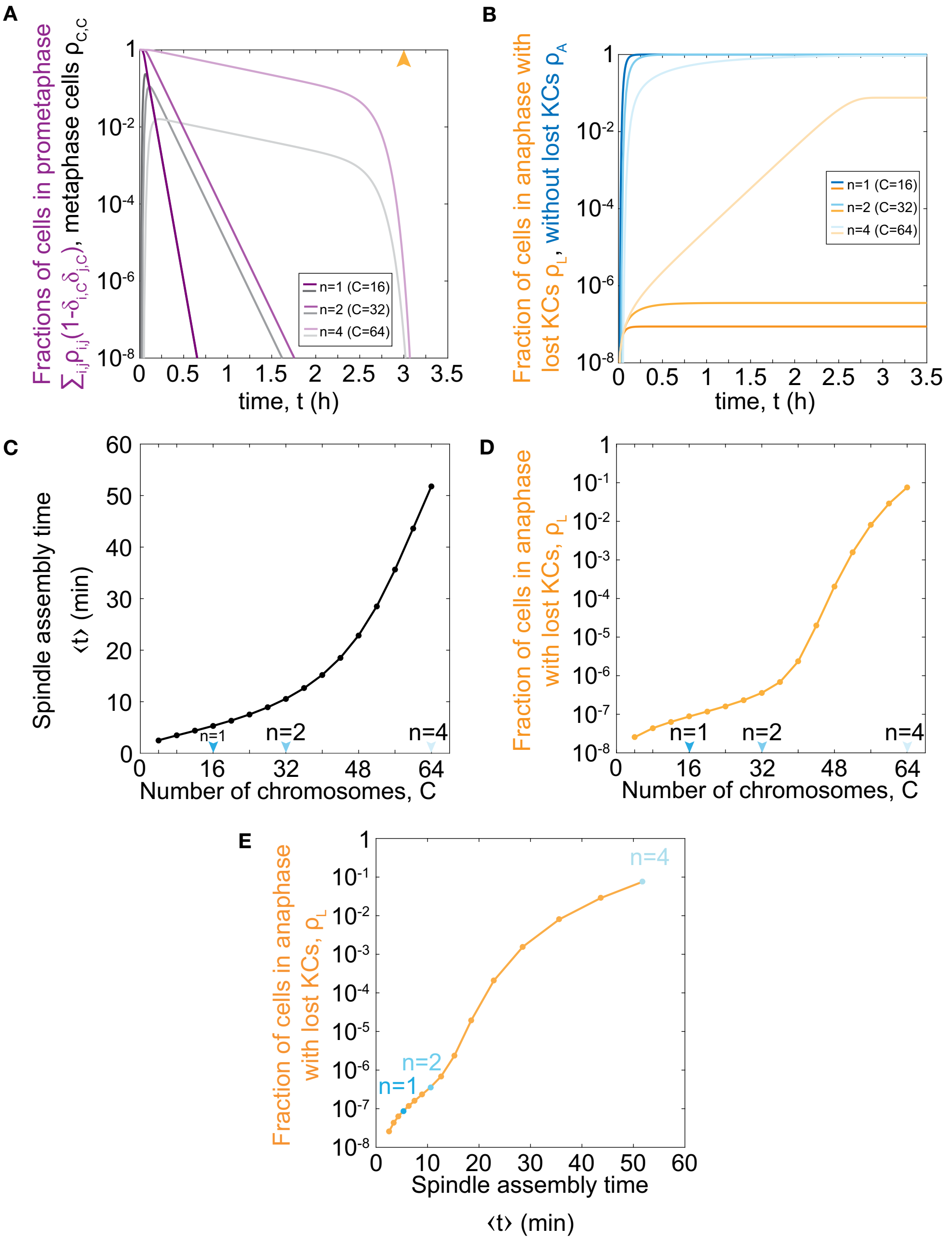


FIGURE 2 | Model predictions for chromosome loss in cells of different ploidy. **(A)** Fraction of cells in prometaphase (purple) and metaphase (gray) for different numbers of chromosomes. The orange arrowhead denotes the value of the duration of mitotic arrest, t_0. **(B)** Fraction of cells in anaphase with at least one KC unattached (orange) and in anaphase (blue). Three different shades in **(A,B)** correspond to different number of chromosomes, $C = 16,\ 32,\ 64$. For color-codes see inset legends. **(C)** Time of spindle assembly as a function of the number of chromosomes. **(D)** Rate of chromosome loss for cells as a function of the number of chromosomes. Arrowheads denote haploid, diploid and tetraploid number of chromosomes. **(E)** Rate of chromosome loss for cells as a function of the time of spindle assembly. Data points are obtained from **(C,D)**, and correspond to $C = 4, \ldots, 64$. Cases with $C = 16,\ 32,\ 64$ are shown in blue. At $t = 0$, $\rho_{0,0} = 1$ and all other populations are 0. The other parameters are given in **Figure 1C**.

and microtubule nucleation sites, α , for different number of chromosomes. For parameter values $\alpha = 1.0$ the time of spindle assembly increases with the number of chromosomes (**Figures 2C**, **3A**). By increasing α to values >1 the assembly speeds up, but the influence is noticeable for a larger number of chromosomes (**Figure 3A**). By decreasing the parameter to the value $\alpha = 0.9$ the assembly time dramatically increases with number of chromosomes and goes to infinity when there are more than 40 chromosomes. The infinite time of spindle assembly occurs for cells in which the number of microtubule

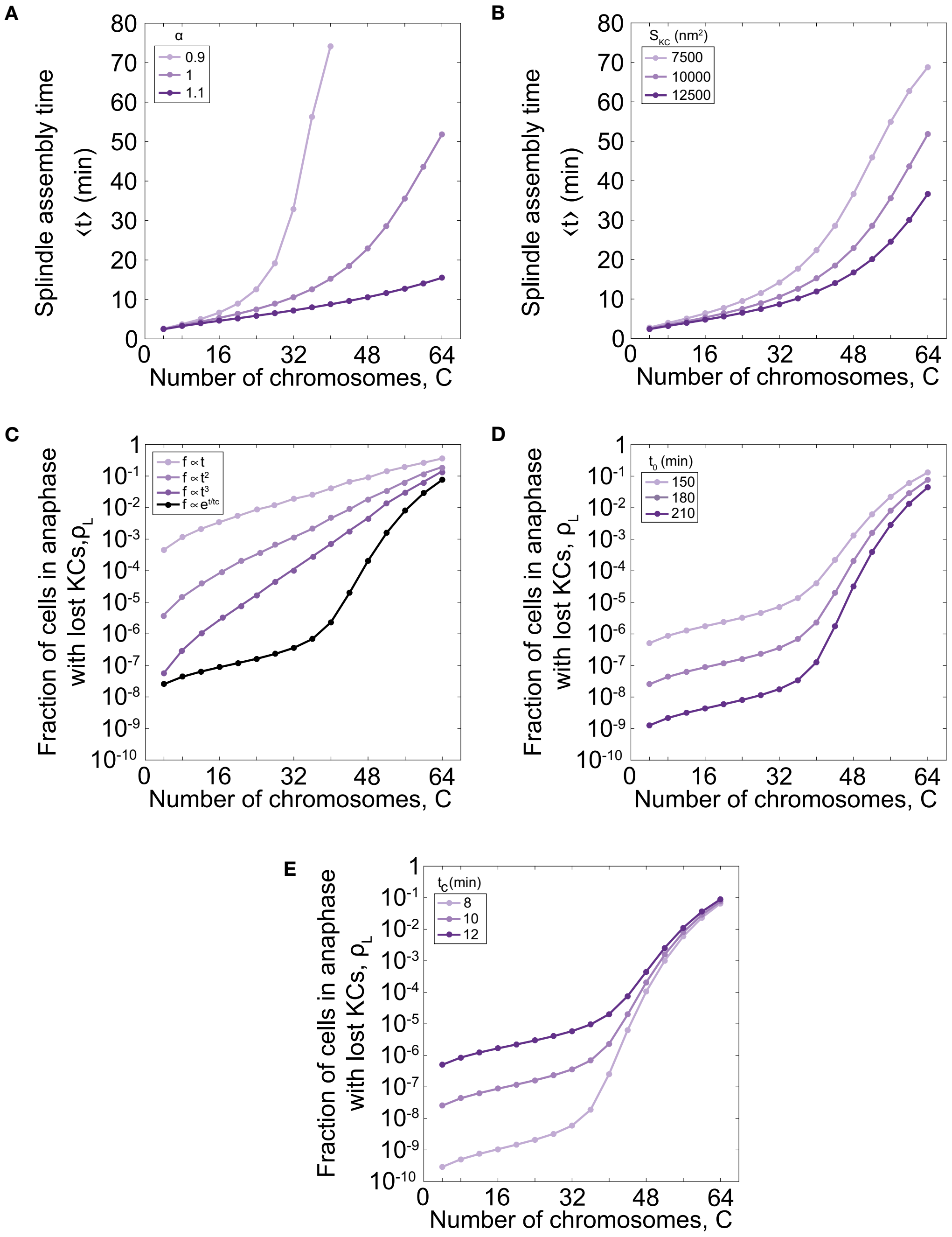


FIGURE 3 | Time of spindle assembly and rate of chromosome loss for different number of chromosomes and different values of model parameters. **(A)** Time of spindle assembly for different number of chromosomes and three different values of $\alpha = 0.9, 1.0, 1.1$. For color-codes see inset legend. The other parameters are given in **Figure 1C**. **(B)** The role of the cross-section area of the KC on the spindle assembly time. Three different shades correspond to different cross-section area of the KC, $S_{KC} = 7500$ nm^2, 10000 nm^2, 12500 nm^2. For color-codes see inset legend. The other parameters are given in **Figure 1C**. **(C)** Rate of chromosome loss for different functional forms of the function $f(t)$: linear function $f = (t_C/t_0^2)t$, quadratic function $f = (t_C/t_0^3)t^2$, cubic function $f = (t_C/t_0^4)t^3$, and exponential function $f = exp\left[(t - t_0)/t_C\right]$. For color-codes see inset legend. The other parameters are given in **Figure 1C**. **(D)** Rate of chromosome loss for different values of the parameter that describe the duration of mitotic arrest, t_0. Three different shades correspond to different values of the parameter $t_0 = 150$ min, 180 min, 210 min. For color-codes see inset legend. The other parameters are given in **Figure 1C**. **(E)** Rate of chromosome loss for different values of the characteristic timescale of mitotic arrest, t_C. Three different shades correspond to different values of the parameter $t_C = 8$ min, 10 min, 12 min. For color-codes see inset legend. The other parameters are given in **Figure 1C**.

nucleation sites at one pole is smaller than number of chromosomes. Interestingly, in yeast the value of the parameter α in cells is close to 1 (**Figure 1C**).

We next explore the relevance of geometry by varying the cross-section area of the KC, S_{KC}. We find that geometry has a small contribution for a small number of chromosomes, but for larger number of chromosomes, the time of spindle assembly decreases with the increase of the cross-section area (**Figure 3B**). The role of the cross-section area occupied by one microtubule, S, can be inferred from these data because both parameters, the cross-section area occupied by one microtubule and the cross-section area of the KC, contribute to attachment probability p.

Further, we explore how the choice of the function that describes bypassing the checkpoint in mitotic arrest $f(t)$ affects the chromosome loss rate. We find that for a linear function the chromosome loss rate increases as the number of chromosome increases (**Figure 3C**). However, in this case the model cannot explain experimental results quantitatively. For example, when number of chromosomes changes from 32 to 64 the chromosome loss rate increases approximately 20 times with the linear function, whereas when ploidy in experiments changes from diploid to tetraploid the loss rate increases thousand times. A chromosome loss rate in the model is more similar to the experimental results for nonlinear functional forms, such as quadratic and cubic functions (**Figure 3C**). Because from this analysis we cannot predict a functional form for the function $f(t)$, we choose an exponential function as a simple function that provides agreement with experiments.

Finally, we explore how the parameters that describe bypassing the checkpoint in mitotic arrest, t_0 and t_c, affect the chromosome loss rate. We find that cells with shorter duration of mitotic arrest have an increased chromosome loss rate, irrespective of ploidy (**Figure 3D**). We also find that cells with a smaller characteristic timescale of mitotic arrest have a smaller rate of chromosome loss (**Figure 3E**).

DISCUSSION

Here we introduced a model in which we explored chromosome loss dynamics by accounting for key aspects of spindle assembly, including microtubule nucleation and KC attachment/detachment, together with a maximum time of mitotic arrest. Our theory provides a plausible explanation for experiments in yeast tetraploid cells, where there is a 1,000-fold increase in the rate of chromosome loss relative to haploid and diploid cells (Mayer and Aguilera, 1990; Storchová et al., 2006). Our model not only quantitatively predicts an increase in chromosome loss in cells with an increasing chromosome number, but also a longer duration of spindle assembly time. Indeed, the doubling time of yeast increases with ploidy in *S. cerevisiae.* For example, doubling times of haploid, diploid and tetraploid yeast cells in YPD is approximately 130, 146, and 171 min, respectively (Mable, 2001). This suggests that cells with increasing ploidy have an increased spindle assembly time, with differences in the same order of magnitude as in our model. However, this prediction needs to be further verified by direct measurements of average spindle assembly time in haploid, diploid, and tetraploid yeast cells. Key parameters of cytoplasmic microtubule dynamics were measured previously for diploid and tetraploid *S. cerevisiae* cells, including the rates of microtubule growth, shrinkage, catastrophe and rescue during G1 and mitosis (Storchová et al., 2006). We hypothesize that changes in these parameters may cause a change in the average spindle assembly time in a population of cells, but experimental validation in yeast is also needed.

In yeast cells of different ploidy, chromosome loss can occur for many reasons. Configurations with syntelic attachments can also appear and lead to chromosome loss. Storchova et al. detected an increased frequency of erroneous KC attachments in polyploid cells and suggest an important role for syntelic attachments based on increased activity of Ipl1, the yeast homolog of Aurora B (Storchová et al., 2006). Additionally, microtubules can detach from KCs during anaphase, which can further increase chromosome loss events. Thus, identifying experimentally which of these configurations are predominant in cells with lost chromosomes is crucial for establishing a complete picture of chromosome loss.

Laboratory tetraploid yeast cells have an increased rate of chromosome loss. However, a recent experimental evolution study with laboratory yeast cells found that some tetraploid cell lines could maintain their full chromosome complement ($C = 64$) for >1,000 generations (Lu et al., 2016). The evolved, stable tetraploid cells had elevated levels of the Sch9 protein, one of the major regulators downstream of TORC1, which is a central regulator of cell growth. Interestingly, the evolved stable tetraploid cells also had increased resistance to the microtubule depolymerizing drug benomyl relative to the ancestor tetraploid cells, indicating that increased Sch9 activity may, at least in part, rescue spindle formation defects observed in the ancestral tetraploid cells (Storchová et al., 2006; Lu et al., 2016). This is consistent with our model, where chromosome stability in tetraploid cells can be obtained by increasing the rate of spindle assembly.

This is the first theoretical study of the mechanism driving high rates of chromosome loss in polyploid yeast cells. Our approach for within-species ploidy variation can be applied to other species, including plants (Hufton and Panopoulou, 2009), where rates of chromosome loss are also higher in polyploid cells than in diploid cells, if the details of spindle self-organization are adjusted for the specific organism and cell-type. For example, for cells with more than one microtubule per KC, merotelic attachments need to be taken into account as well (Gregan et al., 2011). Future models will show the extent to which spindle assembly time influences the rate of chromosome loss for a variety of systems.

AUTHOR CONTRIBUTIONS

NP, LL, and AS conceived the project. NP and LL developed the model, IJ solved the model. All authors wrote the paper.

FUNDING

This research was supported by the QuantiXLie Centre of Excellence, a project co-financed by the Croatian Government and European Union through the European Regional Development Fund – the Competitiveness and Cohesion Operational Programme (Grant KK.01.1.1.01.0004, NP), the Netherlands Organization for Scientific Research (NWO/OCW) as part of the Frontiers of Nanoscience program (LL), a Nebraska EPSCoR First Award (AS) and an LB692-Nebraska Tobacco Settlement Biomedical Research Development Fund (AS).

ACKNOWLEDGMENTS

We thank Timon Idema, Judy Berman and Iva Tolić for critical reading of the manuscript, and Ivana Šarić for the drawings.

REFERENCES

Akiyoshi, B., Sarangapani, K. K., Powers, A. F., Nelson, C. R., Reichow, S. L., Arellano-Santoyo, H., et al. (2010). Tension directly stabilizes reconstituted kinetochore-microtubule attachments. *Nature* 468, 576–579. doi: 10.1038/nature09594

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Fujiwara, T., Bandi, M., Nitta, M., Ivanova, E. V., Bronson, R. T., and Pellman, D. (2005). Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* 437, 1043–1047. doi: 10.1038/nature04217

Ganem, N. J., Godinho, S. A., and Pellman, D. (2009). A mechanism linking extra centrosomes to chromosomal instability. *Nature* 460, 278–282. doi: 10.1038/nature08136

Gardner, M. K., Bouck, D. C., Paliulis, L. V., Meehl, J. B., O'Toole, E. T., Haase, J., et al. (2008). Chromosome congression by kinesin-5 motor-mediated disassembly of longer kinetochore microtubules. *Cell* 135, 894–906. doi: 10.1016/j.cell.2008.09.046

Gay, G., Courtheoux, T., Reyes, C., Tournier, S., and Gachet, Y. (2012). A stochastic model of kinetochore-microtubule attachment accurately describes fission yeast chromosome segregation. *J. Cell Biol.* 196, 757–774. doi: 10.1083/jcb.201107124

Gerstein, A. C., Chun, H. J., Grant, A., and Otto, S. P. (2006). Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet.* 2:e145. doi: 10.1371/journal.pgen.0020145

Gonen, S., Akiyoshi, B., Iadanza, M. G., Shi, D., Duggan, N., Biggins, S., et al. (2012). The structure of purified kinetochores reveals multiple microtubule-attachment sites. *Nat. Struct. Mol. Biol.* 19, 925–929. doi: 10.1038/nsmb.2358

Gregan, J., Polakova, S., Zhang, L., Tolic-Norrelykke, I. M., and Cimini, D. (2011). Merotelic kinetochore attachment: causes and effects. *Trends Cell Biol.* 21, 374–381. doi: 10.1016/j.tcb.2011.01.003

Hill, T. L. (1985). Theoretical problems related to the attachment of microtubules to kinetochores. *Proc. Natl. Acad. Sci. U.S.A.* 82, 4404–4408. doi: 10.1073/pnas.82.13.4404

Hufton, A. L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.* 19, 600–606. doi: 10.1016/j.gde.2009.10.005

Kalinina, I., Nandi, A., Delivani, P., Chacon, M. R., Klemm, A. H., Ramunno-Johnson, D., et al. (2013). Pivoting of microtubules around the spindle pole accelerates kinetochore capture. *Nat. Cell Biol.* 15, 82–87. doi: 10.1038/ncb2640

Kitamura, E., Tanaka, K., Komoto, S., Kitamura, Y., Antony, C., and Tanaka, T. U. (2010). Kinetochores generate microtubules with distal plus ends: their roles and limited lifetime in mitosis. *Dev. Cell* 18, 248–259. doi: 10.1016/j.devcel.2009.12.018

Lea, D. E., and Coulson, C. A. (1949). The distribution of the numbers of mutants in bacterial populations. *J. Genet.* 49, 264–285. doi: 10.1007/BF02986080

Li, R., and Murray, A. W. (1991). Feedback control of mitosis in budding yeast. *Cell* 66, 519–531. doi: 10.1016/0092-8674(81)90015-5

Lu, Y. J., Swamy, K. B., and Leu, J. Y. (2016). Experimental evolution reveals interplay between Sch9 and polyploid stability in yeast. *PLoS Genet.* 12:e1006409. doi: 10.1371/journal.pgen.1006409

Mable, B. K. (2001). Ploidy evolution in the yeast *Saccharomyces cerevisiae*: a test of the nutrient limitation hypothesis. *J. Evol. Biol.* 14, 157–170. doi: 10.1046/j.1420-9101.2001.00245.x

Mayer, V. W., and Aguilera, A. (1990). High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat. Res.* 231, 177–186. doi: 10.1016/0027-5107(90)90024-X

Minshull, J., Straight, A., Rudner, A. D., Dernburg, A. F., Belmont, A., and Murray, A. W. (1996). Protein phosphatase 2A regulates MPF activity and sister chromatid cohesion in budding yeast. *Curr. Biol.* 6, 1609–1620. doi: 10.1016/S0960-9822(02)70784-7

Mitchison, T. J., and Kirschner, M. W. (1985). Properties of the kinetochore *in vitro*. II. Microtubule capture and ATP-dependent translocation. *J. Cell. Biol.* 101, 766–777. doi: 10.1083/jcb.101.3.766

Musacchio, A., and Salmon, E. D. (2007). The spindle-assembly checkpoint in space and time. *Nat. Rev. Mol. Cell Biol.* 8, 379–393. doi: 10.1038/nrm2163

Nannas, N. J., O'Toole, E. T., Winey, M., and Murray, A. W. (2014). Chromosomal attachments set length and microtubule number in the *Saccharomyces cerevisiae* mitotic spindle. *Mol. Biol. Cell* 25, 4034–4048. doi: 10.1091/mbc.e14-01-0016

Novák, B., Tóth, A., Csikász-Nagy, A., Gyorffy, B., Tyson, J. J., and Nasmyth, K. (1999). Finishing the cell cycle. *J. Theor. Biol.* 199, 223–233. doi: 10.1006/jtbi.1999.0956

O'Toole, E. T., Mastronarde, D. N., Giddings, T. H. Jr., Winey, M., Burke, D. J., and McIntosh, J. R. (1997). Three-dimensional analysis and ultrastructural design of mitotic spindles from the cdc20 mutant of *Saccharomyces cerevisiae*. *Mol. Biol. Cell* 8, 1–11. doi: 10.1091/mbc.8.1.1

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437. doi: 10.1146/annurev.genet.34.1.401

Paul, R., Wollman, R., Silkworth, W. T., Nardi, I. K., Cimini, D., and Mogilner, A. (2009). Computer simulations predict that chromosome movements and rotations accelerate mitotic spindle assembly without compromising accuracy. *Proc. Natl. Acad. Sci. U.S.A.* 106, 15708–15713. doi: 10.1073/pnas.0908261106

Pavin, N., and Tolić, I. M. (2016). Self-organization and forces in the mitotic spindle. *Annu. Rev. Biophys.* 45, 279–298. doi: 10.1146/annurev-biophys-062215-010934

Prosser, S. L., and Pelletier, L. (2017). Mitotic spindle assembly in animal cells: a fine balancing act. *Nat. Rev. Mole. Cell Biol.* 18, 187–201. doi: 10.1038/nrm.2016.162

Raina, S. N., Parida, A., Koul, K. K., Salimath, S. S., Bisht, M. S., Raja, V., et al. (1994). Associated chromosomal DNA changes in polyploids. *Genome* 37, 560–564. doi: 10.1139/g94-080

Rieder, C. L., and Maiato, H. (2004). Stuck in division or passing through: what happens when cells cannot satisfy the spindle assembly checkpoint. *Dev. Cell* 7, 637–651. doi: 10.1016/j.devcel.2004.09.002

Rudner, A. D., Hardwick, K. G., and Murray, A. W. (2000). Cdc28 activates exit from mitosis in budding yeast. *J. Cell. Biol.* 149, 1361–1376. doi: 10.1083/jcb.149.7.1361

Rudner, A. D., and Murray, A. W. (1996). The spindle assembly checkpoint. *Curr. Opin. Cell Biol.* 8, 773–780. doi: 10.1016/S0955-0674(96)80077-9

Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., et al. (2015). Polyploidy can drive rapid adaptation in yeast. *Nature* 519, 349–352. doi: 10.1038/nature14187

Song, K., Lu, P., Tang, K., and Osborn, T. C. (1995). Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proc. Natl. Acad. Sci. U.S.A.* 92, 7719–7723. doi: 10.1073/pnas.92.17.7719

Storchová, Z., Breneman, A., Cande, J., Dunn, J., Burbank, K., O'Toole, E., et al. (2006). Genome-wide genetic analysis of polyploidy in yeast. *Nature* 443, 541–547. doi: 10.1038/nature05178

Tanaka, K., Mukae, N., Dewar, H., van Breugel, M., James, E. K., Prescott, A. R., et al. (2005). Molecular mechanisms of kinetochore capture by spindle microtubules. *Nature* 434, 987–994. doi: 10.1038/nature03483

Tubman, E. S., Biggins, S., and Odde, D. J. (2017). Stochastic modeling yields a mechanistic framework for spindle attachment error correction in budding yeast mitosis. *Cell. Syst.* 4, 645–650.e5, doi: 10.1016/j.cels.2017.05.003

Vasileva, V., Gierlinski, M., Yue, Z., O'Reilly, N., Kitamura, E., and Tanaka, T. U. (2017). Molecular mechanisms facilitating the initial kinetochore encounter with spindle microtubules. *J. Cell. Biol.* 216, 1609–1622. doi: 10.1083/jcb.201608122

Volkov, V. A., Zaytsev, A. V., Gudimchuk, N., Grissom, P. M., Gintsburg, A. L., Ataullakhanov, F. I., et al. (2013). Long tethers provide high-force coupling of the Dam1 ring to shortening microtubules. *Proc. Natl. Acad. Sci. U.S.A.* 110, 7708–7713. doi: 10.1073/pnas.1305821110

Winey, M., Mamay, C. L., O'Toole, E. T., Mastronarde, D. N., Giddings, T. H. Jr., McDonald, K. L., et al. (1995). Three-dimensional ultrastructural analysis of the *Saccharomyces cerevisiae* mitotic spindle. *J. Cell Biol.* 129, 1601–1615. doi: 10.1083/jcb.129.6.1601

Wollman, R., Cytrynbaum, E. N., Jones, J. T., Meyer, T., Scholey, J. M., and Mogilner, A. (2005). Efficient chromosome capture requires a bias in the 'search-and-capture' process during mitotic-spindle assembly. *Curr. Biol.* 15, 828–832. doi: 10.1016/j.cub.2005.03.019

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* 45, 1134–1140. doi: 10.1038/ng.2760

Zaytsev, A. V., and Grishchuk, E. L. (2015). Basic mechanism for biorientation of mitotic chromosomes is provided by the kinetochore geometry and indiscriminate turnover of kinetochore microtubules. *Mol. Biol. Cell* 26, 3985–3998. doi: 10.1091/mbc.e15-06-0384

Conflict of Interest Statement: 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 Plant Science

ORIGINAL RESEARCH
published: 08 August 2018
doi: 10.3389/fpls.2018.01163

Gene Co-expression Network Analysis Suggests the Existence of Transcriptional Modules Containing a High Proportion of Transcriptionally Differentiated Homoeologs in Hexaploid Wheat

Kotaro Takahagi[1,2,3], *Komaki Inoue*[1] *and Keiichi Mochida*[1,2,3,4]*

[1] Bioproductivity Informatics Research Team, RIKEN Center for Sustainable Resource Science, Yokohama, Japan, [2] Graduate School of Nanobioscience, Yokohama City University, Yokohama, Japan, [3] Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan, [4] Institute of Plant Science and Resources, Okayama University, Kurashiki, Japan

OPEN ACCESS

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Divya Mehta,
Queensland University of Technology, Australia
Wilco Ligterink,
Wageningen University & Research, Netherlands

***Correspondence:**
Keiichi Mochida
keiichi.mochida@riken.jp

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

Received: *31 January 2018*
Accepted: *23 July 2018*
Published: *08 August 2018*

Citation:
Takahagi K, Inoue K and Mochida K (2018) Gene Co-expression Network Analysis Suggests the Existence of Transcriptional Modules Containing a High Proportion of Transcriptionally Differentiated Homoeologs in Hexaploid Wheat. Front. Plant Sci. 9:1163. doi: 10.3389/fpls.2018.01163

Genome duplications aid in the formation of novel molecular networks through regulatory differentiation of the duplicated genes and facilitate adaptation to environmental change. Hexaploid wheat, *Triticum aestivum*, contains three homoeologous chromosome sets, the A-, B-, and D-subgenomes, which evolved through interspecific hybridization and subsequent whole-genome duplication. The divergent expression patterns of the homoeologs in hexaploid wheat suggest that they have undergone transcriptional and/or functional differentiation during wheat evolution. However, the distribution of transcriptionally differentiated homoeologs in gene regulatory networks and their related biological functions in hexaploid wheat are still largely unexplored. Therefore, we retrieved 727 publicly available wheat RNA-sequencing (RNA-seq) datasets from various tissues, developmental stages, and conditions, and identified 10,415 expressed homoeologous triplets. Examining the co-expression modules in the wheat transcriptome, we found that 66% of the expressed homoeologous triplets possess all three homoeologs grouped in the same co-expression modules. Among these, 15 triplets contain co-expressed homoeologs with differential expression levels between homoeoalleles across $\geq$ 95% of the 727 RNA-seq datasets, suggesting a consistent trend of homoeolog expression bias. In addition, we identified 2,831 differentiated homoeologs that showed gene expression patterns that deviated from those of the other two homoeologs. We found that seven co-expression modules contained a high proportion of such differentiated homoeologs, which accounted for $\geq$ 20% of the genes in each module. We also found that five of the co-expression modules are abundantly composed of genes involved in biological processes such as chloroplast biogenesis, RNA metabolism, putative defense response, putative posttranscriptional modification, and lipid metabolism, thereby suggesting that, the differentiated homoeologs might highly contribute to these biological functions in the gene network of hexaploid wheat.

Keywords: allopolyploidization, co-expression gene network, hexaploid wheat, homoeolog, transcriptional module

INTRODUCTION

Interspecific hybridization and polyploidization have played important roles in the evolution and diversification of plants (Soltis and Soltis, 2009; Van de Peer et al., 2009). Allopolyploids are originated from hybridization between different species followed by whole-genome duplication (Ramsey and Schemske, 1998; Comai, 2005). Despite the multiple conditions that need to be met for allopolyploidization to occur, including existing populations of parental lines in the same area, overcoming hybrid incompatibility, gametic non-reduction, and chromosome doubling (Osabe et al., 2012), the occurrence of allopolyploids is widespread in various taxonomic groups in plants (Leitch and Leitch, 2008; Barker et al., 2016). Therefore, it has been hypothesized that allopolyploid species have evolutionary advantages compared to their diploid ancestral species (Wendel, 2000; Doyle et al., 2008).

Improved traits that evolved in allopolyploid plants enhanced their productivity and have contributed to the domestication of many crops (Chen, 2010; Renny-Byfield and Wendel, 2014). For example, the allotetraploid *Arabidopsis suecica* has more vigorous growth and produces more seeds than its ancestral species (Solhaug et al., 2016), whereas the allotetraploid *Coffea arabica* can better adapt to changes in temperature than its diploid ancestors (Combes et al., 2013). In allohexaploid wheat (*Triticum aestivum*), both natural and synthetic plants have higher tolerance to salt stress than their diploid and tetraploid ancestors (Dubcovsky and Dvorak, 2007; Yang et al., 2014). These examples suggest that allopolyploidization often leads to increased productivity through fixation of genomic heterozygosity, which improves environmental fitness and contributes to the habitat expansion of a species.

Allopolyploidization can give rise to transcriptional and/or functional changes in homoeologs (genes that are duplicated due to allopolyploidization) (Mochida et al., 2003; Adams and Wendel, 2005; Moore and Purugganan, 2005). Homoeologs can undergo accelerated evolution due to redundant genetic codes that can evolve new functions without constraints (Kaessmann, 2010; Naseeb et al., 2017). A number of studies have revealed their fates as non-functionalized (loss of function of one of the duplicated genes), subfunctionalized (partitioning of function between duplicated genes), and/or neo-functionalized (diversification of function between the duplicated genes) (Lynch and Conery, 2000; Blanc and Wolfe, 2004; Cusack and Wolfe, 2007). Homoeologs in plants often show different expression patterns across tissues, developmental stages, and conditions, suggesting that they have undergone sub- and/or neofunctionalization (Madlung, 2013). The differential employment of homoeologs through dynamic transcriptional regulation may contribute to the enhanced evolutionarily adaptability of allopolyploid species.

A number of studies based on homoeolog-specific gene expression analysis have reported the evolutionary fates of homoeologs in various allopolyploid plants (Adams, 2007; Hughes et al., 2014; Takahagi et al., 2018). Transcriptome analysis has revealed that the expression of multiple ribosomal protein-coding homoeologs in *Brassica napus* is tissue-dependent (Whittle and Krochko, 2009). An investigation of the relative levels of allelic and homoeologous gene expression in cotton revealed that subfunctionalized genes are mainly expressed in reproductive tissues, and non-functionalized alleles are typically derived from the A-genome, indicating potential genome-of-origin bias for neofunctionalization (Chaudhary et al., 2009). Differentiation of expression patterns of homoeologs in allopolyploid species might effect changes in their gene regulatory networks owing to transcriptional and/or functional divergence. The evolutionary changes in gene regulatory networks are thought to facilitate responses to developmental programs and environmental cues in allopolyploids (Chen and Ni, 2006).

Hexaploid wheat, *Triticum aestivum*, is a widely cultivated allohexaploid crop (2n = 6x = 42, AABBDD) that originated from hybridization between the domesticated allotetraploid *Triticum turgidum* (2n = 4x = 28, AABB) and the diploid goat grass *Aegilops tauschii* (2n = 2x = 14, DD) approximately 10,000 years ago, followed by genome duplication (Matsuoka, 2011; Feldman and Levy, 2012). Pfeifer et al. (2014) generated a co-expression gene network of hexaploid wheat and examined the contribution of expression of each homoeolog. They found that several network modules exhibit unbalanced homoeolog expression, which might be associated with biological functions and tissue types (Pfeifer et al., 2014). Recently, Tanaka et al. (2016) reported homoeolog-specific regulation of the floral MADS-box genes in wheat, and differential expression patterns of homoeologs were consistently observed in both natural and synthetic allohexaploid wheat varieties (Tanaka et al., 2016). Moreover, Powell et al. (2017) demonstrated that the wheat transcriptome has homoeolog expression bias toward the B- and D-subgenomes in response to pathogen infection (Powell et al., 2017). The divergent expression patterns between homoeologs suggest that they have undergone transcriptional and/or functional differentiation. However, the distribution of transcriptionally differentiated homoeologs in gene regulatory networks and their related biological functions in hexaploid wheat are still largely unexplored.

In this study, to elucidate homoeologous networks in hexaploid wheat and to explore their differentiation, we retrieved publicly available RNA-sequencing (RNA-seq) datasets from various tissues, developmental stages, and conditions. We categorized hexaploid wheat genes to construct homoeologous groups and identified expressed homoeologous triplets. We also identified differentiated homoeologs that show gene expression patterns that deviate from those of the other two homoeologs. In addition, we explored gene network modules containing a high proportion of differentiated homoeologs in the transcriptome of hexaploid wheat. We assessed enriched functions in the network modules and discussed the evolution of such network modules resulting from transcriptional differentiation of homoeologs in hexaploid wheat.

MATERIALS AND METHODS

Data and Data Processing

All publicly available wheat transcriptome sequence datasets were retrieved from the NCBI Sequence Read Archive (April 26, 2017)[1]. To adjust the data format, the datasets were screened according to the following criteria: (1) RNA-seq data strictly (i.e., no EST, FL-cDNA, etc.) from *Triticum aestivum* samples, (2) total number of sequence reads ≥ 10,000,000, and (3) an average sequence read length is 70–1000 bases. The RNA-seq datasets presenting the following characteristics were also removed from analyses, as they were considered inappropriate for gene expression profiling: (1) datasets resulting from pooled samples, taken at different time points, (2) datasets obtained from chromosome deletion and chromosome addition lines, and (3) datasets obtained for poorly described methodologies. RNA-seq reads of the screened datasets were trimmed using Trimmomatic (v.0.32) (Bolger et al., 2014) with the following settings: -thread 1 LEADING: 20 TRAILING: 20 SLIDINGWINDOW:4:15 MINLEN: 50. To obtain high-quality sequence datasets, the trimmed datasets were further screened according to the following criteria: (1) ≥ 70% of raw reads are maintained after the trimming step and (2) an average sequence read length is 70–1000 bases after trimming. The trimmed reads obtained after the second screening were mapped to the representative cDNA sequences annotated in the genome assembly of Chinese Spring wheat (International Wheat Genome Sequencing Consortium, 2014) downloaded from the Ensembl (v.35)[2] using the BWA program (v.0.7.8) (Li and Durbin, 2009) with its mem command. To use datasets with high-quality alignments of the reads, those that were not uniquely mapped and/or not paired mapped were removed from the read alignment datasets using custom Perl scripts. In total, 727 read alignment datasets (**Supplementary Table S1**), for which ≥ 50% of raw reads remained after the read removal step, were subjected to further analysis. The reads per million mapped reads (RPM) values were calculated for all genes in the 727 read alignment datasets. Genes with an RPM ≥ 3 in at least eight datasets (≥ 1% of the 727 RNA-seq datasets) were identified as significantly expressed genes.

Identification of Homoeologous Groups

To identify homoeologous groups, representative protein sequences of the A-, B-, and D-subgenomes annotated in the genome assembly of Chinese Spring wheat (International Wheat Genome Sequencing Consortium, 2014) downloaded from Ensembl (v.35)[2] were compared against each other using BLASTP (v.2.6) (McGinnis and Madden, 2004), applying an *e*-value cut-off of 1e-5 and a sequence identity cut-off of 90%. Sets of three homoeologs that were reciprocal best hits in all pairwise comparisons were identified as homoeologous triplets (ABD type in **Figure 1B**). Sets of two homoeologs with reciprocal best hits for two subgenomes and without hits for the other subgenome were identified as homoeologous doublets (AB, AD, and BD types in **Figure 1B**). Genes without hits in any of the other two subgenomes were identified as subgenome-unique genes (A, B, and D types in **Figure 1B**).

t-distributed Stochastic Neighbor Embedding (t-SNE) Analysis

To summarize expression patterns of the genes with an RPM ≥ 3 in a range of 1–7 datasets (spatiotemporally expressed genes), t-SNE analysis was performed using the Rtsne package (v.0.13)[3] in R (v.3.4.3). The number of iterations was set at 10,000, and parameter theta was set at 0.0.

Co-expression Network Analysis

To compute co-expression modules of homoeologs, WGCNA analysis (Langfelder and Horvath, 2008) was performed based on the normalized RPM using the one-step automatic network construction method with the following parameters: power = 9, networkType = "signed", TOMType = "unsigned", minModuleSize = 30, reassignThreshold = 0, mergeCutHeight = 0.25, numericLabels = TRUE, pamRespectsDendro = FALSE. A soft-thresholding power was selected by evaluating the scale-free topology model fit.

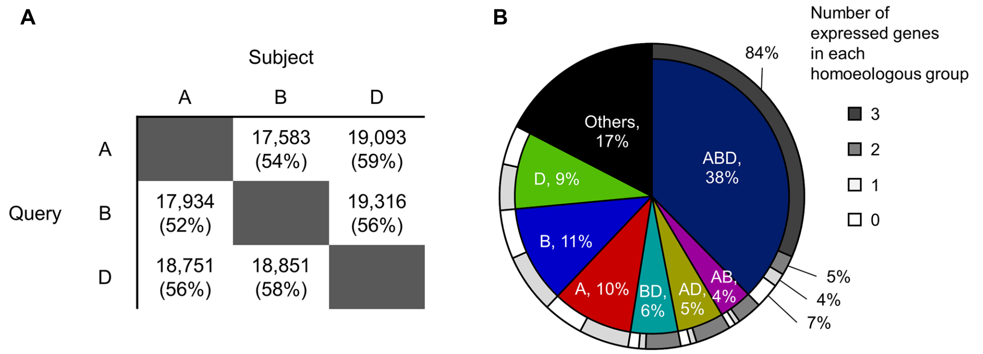
Identification of Differentially Expressed Genes

For identification of the homoeologous triplets containing co-expressed homoeologs with differential expression levels between homoeoalleles, the gene expression fold changes between homoeologs across the 727 RNA-seq datasets were calculated based on RPM. Pairs of homoeologs with a fold change ≥ 3 and RPM ≥ 3 for at least one of the homoeologs were identified as differentially expressed homoeologs. For the examination of expression bias between homoeologs in the homoeologous triplets, reads used for RPM calculation in a series of RNA-seq datasets (SRR1542404-SRR1542417) (Liu et al., 2015) were subjected to differential gene expression analysis performed by using the edgeR package (v.3.20.9) (Robinson et al., 2010) in R (v.3.4.3). Pairs of homoeologs with a false discovery rate (FDR) ≤ 0.001 and RPM (average of 2 biological replicates in the

[1] https://www.ncbi.nlm.nih.gov/sra

[2] http://plants.ensembl.org/Triticum_aestivum/Info/Index

[3] https://github.com/jkrijthe/Rtsne



Query	A	B	D
A		17,583 (54%)	19,093 (59%)
B	17,934 (52%)		19,316 (56%)
D	18,751 (56%)	18,851 (58%)	

FIGURE 1 | Homoeologous groups in hexaploid wheat. **(A)** Numbers of A-, B-, and D-homoeologs that show high sequence similarity with the other two subgenomes based on BLAST analysis. The e-value cut-off was set at 1e-5 and the sequence identity cut-off was set to 90%. Values in brackets are percentages of the total number of query sequences. **(B)** Proportions of genes classified into each homoeologous group. ABD: sets of three homoeologs that are reciprocal best hits in all pairwise comparisons (i.e., homoeologous triplet); AB, AD, and BD: sets of two homoeologs with reciprocal best hits for two subgenomes and without hits for the other subgenome (i.e., homoeologous doublets); A, B, and D: genes without hits in any of the other two subgenomes (i.e., subgenome-unique genes); Others: genes that are not clustered into an homoeologous groups (e.g., genes with BLAST hits for the other subgenome(s) but that are not reciprocal best hits). The outer circle shows proportions of the number of expressed genes in each homoeologous group.

RNA-seq datasets) $\geq$ 3 for at least one of the homoeologs were identified as significantly differentially expressed homoeologs.

Gene Ontology (GO) Enrichment Analysis

The closest homologs of wheat genes in Arabidopsis and rice were identified by BLASTP (v.2.6) (McGinnis and Madden, 2004) searches, applying an *e*-value threshold of $\leq$ 1e-5. GO terms of the best-hit genes in Arabidopsis and rice were used as the customized annotations for wheat genes. To reduce bias, GO terms that were assigned to more than 5,000 wheat genes were excluded. Enriched GO terms were identified for selected genes using BLAST2GO (v.4.1.9) (Conesa et al., 2005) with the customized annotations of wheat genes. For the estimation of the enriched GO terms of genes that are spatiotemporally expressed (representing genes with an RPM $\geq$ 3 in less than 1% (eight datasets) of the 727 RNA-seq datasets) or non-significantly expressed (representing genes with an RPM < 3 in all of the 727 RNA-seq datasets), all of the annotated genes in the Chinese Spring wheat chromosomes were used as a reference set. For estimation of the enriched GO terms of the other sets of genes, those in the expressed homoeologous triplets were used as a reference set. The significance threshold was set at FDR $\leq$ 0.001. The enriched GO terms were summarized based on their semantic similarities using the web-based tool REVIGO[4] (Supek et al., 2011).

RESULTS

Homoeologous Triplets in Hexaploid Wheat

To explore the distribution of transcriptionally differentiated homoeologs in gene regulatory networks and their related biological functions in hexaploid wheat, we identified expressed homoeologous triplets using publicly available RNA-seq datasets. We gathered 727 RNA-seq datasets from hexaploid wheat composed of as many as 517 biosamples relating to various tissues, developmental stages, and conditions, which enabled us to comprehensively explore functional differentiation of transcription regulatory networks in hexaploid wheat (**Supplementary Table S1**). We mapped the quality-checked reads of the RNA-seq datasets to the set of representative cDNA sequences annotated in the genome assembly of Chinese Spring wheat. Using a threshold of RPM $\geq$ 3 in at least eight datasets ($\geq$1% of the 727 RNA-seq datasets), we found that 73,329 genes (74% of the 99,308 genes corresponding to the representative cDNA sequences assigned to each chromosome) are significantly expressed in hexaploid wheat. To construct putative homoeologous groups, and estimate the number of expressed homoeologs from each homoeoloci, we clustered all the 99,308 genes into 49,710 gene groups based on sequence similarity, using a reciprocal BLAST homology search (**Figure 1A**). Approximately 38% of the genes were classified into gene groups composed of three homoeologs, one from each subgenome (homoeologous triplets, ABD type in **Figure 1B**), in which 84% of the triplets (10,415 triplets) contained three homoeologs significantly expressed in the RNA-seq datasets (expressed homoeologous triplets; **Figure 1B**). We also observed that 31,738 genes (39% of 82,012 genes assigned into each of the homoeologous groups) are expressed from one or two homoeologous loci on the subgenomes, which suggests that approximately 40% of the homoeologous groups contain homoeologs rarely expressed or silenced in the wheat transcriptome (**Figure 1B**).

Spatiotemporally Expressed Genes in Wheat

To characterize the genes found in the wheat transcriptome that are rarely expressed or silenced, we investigated the

[4] http://revigo.irb.hr/

chromosomal distribution and function of these genes. Using the threshold to identify significantly expressed genes, we classified 25,979 genes as rarely expressed or silenced, which suggested a transcriptional sign of non-functionalization or acceleration of spatiotemporal transcriptional regulation. To further investigate the functional properties of such genes, we assessed their chromosomal distribution; however, no biased distribution of these genes was found across the 21 wheat chromosomes (**Figure 2A**). We found that 44% of the 25,979 genes were expressed in at least one RNA-seq dataset with an RPM ≥ 3, whereas the remaining 56% genes showed an RPM < 3 in all of the RNA-seq datasets, suggesting spatiotemporal expression and insignificant expression, respectively (**Figure 2B**). To summarize the expression patterns of the spatiotemporally expressed genes across the 727 RNA-seq datasets, we clustered and visualized the expression profiles of these genes using the t-SNE algorithm, and detected several clusters corresponding to the RNA-seq datasets from particular tissues, such as roots, stamens, and anthers (**Figure 2C**), suggesting their tissue-specific expressions. To assess gene functions over-represented in the spatiotemporally or non-significantly expressed genes, we performed GO enrichment analysis, and found some enriched GO terms related to the response to abiotic stresses, metabolism, and organ development (**Figure 2D**).

Expression Bias Between Homoeologs in Hexaploid Wheat

To examine expression bias between homoeologs in the expressed homoeologous triplets, we computed co-expressed homoeologs and differentially expressed homoeologs based on the 727 RNA-seq datasets. For identification of the co-expressed homoeologs, we applied the WGCNA algorithm, and identified 22 co-expression modules. The results of WGCNA analysis indicated that 66% of the expressed homoeologous triplets possess all three homoeologs grouped in the same co-expression modules (co-expressed triplets, ABD type in **Figure 3A**). For 27% of the triplets, two out of three homoeologs were grouped in the same co-expression modules (AB-D, AD-B, and BD-A types in **Figure 3A**), whereas for the remaining 5% of the triplets, all three homoeologs were assigned to different modules (A-B-D type in **Figure 3A**). To further identify homoeologs that are co-expressed while differentially expressed (representing similar expression patterns across the 727 RNA-seq datasets and differential expression levels between homoeoalleles), we identified differentially expressed homoeologs (fold-change ≥ 3) in the co-expressed triplets, and found that at least 258 triplets contained co-expressed homoeologs with differential expression levels between homoeoalleles across $\geq$ 50% of the 727 RNA-seq datasets (**Figure 3B**). We also found that 15 co-expressed triplets contained such homoeologs observed in $\geq$ 95% of the datasets, suggesting a consistent trend of homoeolog expression bias (**Figures 3B,C**). On the basis of our GO enrichment analysis of these genes, we observed several over-represented functions, such as biotin metabolism, protein modifications, and response to gibberellin stimulus (**Figure 3D**). Moreover, to illuminate homoeolog-specific expression patterns relative to particular tissue type that are supported statistically, we examined the expression bias between homoeologs in the homoeologous triplets in a series of RNA-seq datasets related to multiple abiotic stress conditions such as drought, heat, and combined heat and drought (SRR1542404-SRR1542417) (Liu et al., 2015), and found that an increased number of homoeologous triplets contained differentially expressed homoeologs (FDR $\leq$ 0.001) in response to the drought and heat stress conditions, thereby suggesting the differentiation of transcriptional responsiveness between homoeologs to environmental stresses (**Supplementary Table S2**).

Transcriptional Modules Containing a Number of Differentiated Homoeologs

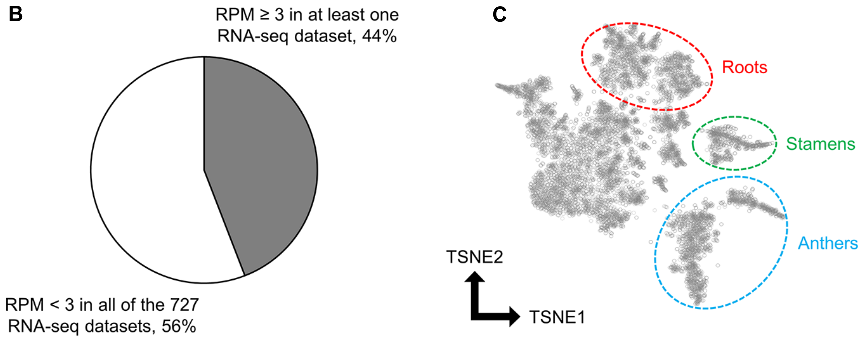
We constructed co-expression gene networks based on the 727 RNA-seq datasets, and thus found that differentiated homoeologs were unevenly distributed in each of the co-expression modules and that several modules contained high proportions of differentiated homoeologs. On the basis of co-expression modules established from our WGCNA analysis, we identified 2,831 homoeologous triplets containing one homoeolog for which the expression pattern deviated from those of the other two homoeologs, which consisted of 9, 10, and 8% of differentiated homoeologs located in A-, B-, and D-subgenomes, respectively (BD-A, AD-B, and AB-D types, respectively, in **Figure 3A**). We also found that such differentiated homoeologs accounted for approximately 9% of all genes used for the WGCNA analysis (10,415 homoeologous triplets; 31,245 genes), whereas seven co-expression modules contained a high proportion of differentiated homoeologs, accounting for $\geq$ 20% of the genes in each module (**Figure 4A**). To estimate enriched biological functions for the genes within the co-expression modules containing a number of differentiated homoeologs, we performed GO enrichment analysis, and found that five of the co-expression modules are abundant in genes involved in biological processes such as chloroplast biogenesis (module 7; **Figure 4B**), RNA metabolism (module 8; **Figure 4C**), putative defense response (module 10; **Figure 4D**), putative posttranscriptional modification (module 15; **Figure 4E**), and lipid metabolism (module 18; **Figure 4F**). These findings suggest that differentiated homoeologs might highly contribute to these biological functions in the gene network of hexaploid wheat.

DISCUSSION

Through our homoeologous gene expression analysis of hexaploid wheat based on a number of RNA-seq datasets, we demonstrated a landscape of transcriptional differentiation among homoeologs. Our comprehensive list of genes that were significantly expressed from one or two homoeologous loci enabled us to identify those genes that may have undergone transcriptional suppression or be directed to spatiotemporal expression. Leach et al. (2014) reported that 55% of genes in hexaploid wheat are expressed from one or two homoeologous loci on the subgenomes in root and shoot tissues (Leach et al., 2014). Using the RNA-seq datasets of 90 wheat lines,

A

	Chromosome group 1	Chromosome group 2	Chromosome group 3	Chromosome group 4	Chromosome group 5	Chromosome group 6	Chromosome group 7
A	1,003 (3.9%)	1,417 (5.5%)	1,187 (4.6%)	1,098 (4.2%)	1,276 (4.9%)	984 (3.8%)	1,370 (5.3%)
B	1,144 (4.4%)	1,685 (6.5%)	1,590 (6.1%)	830 (3.2%)	1,456 (5.6%)	1,281 (4.9%)	1,359 (5.2%)
D	989 (3.8%)	1,533 (5.9%)	1,300 (5.0%)	789 (3.0%)	1,319 (5.1%)	898 (3.5%)	1,471 (5.7%)



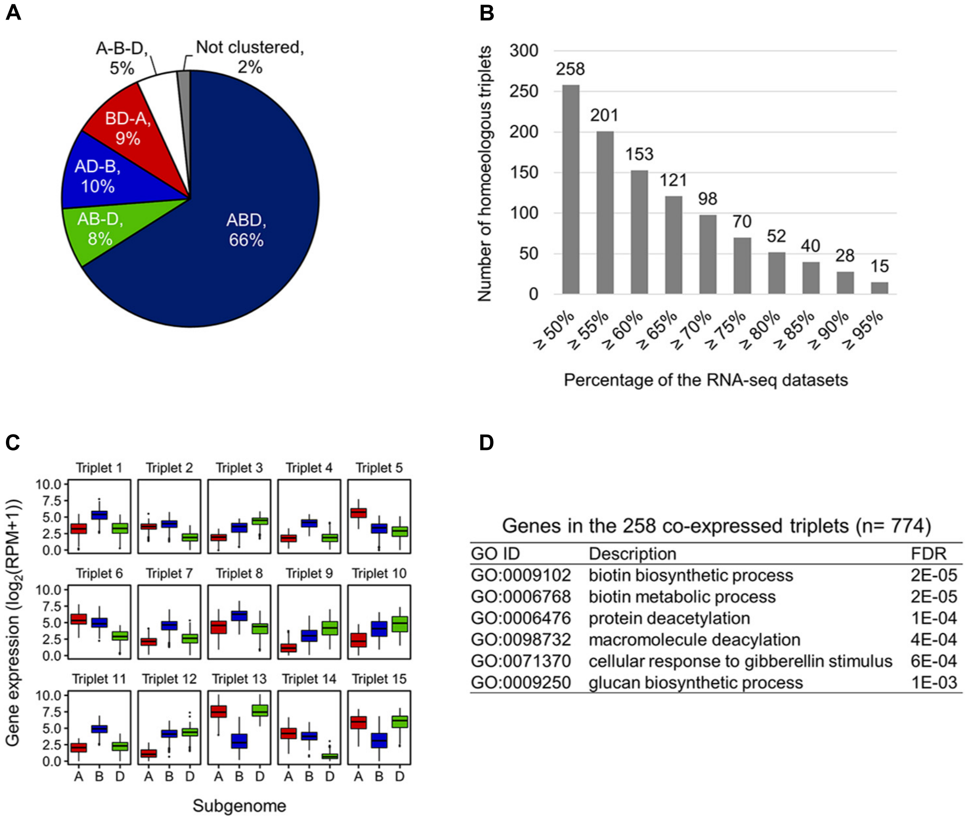
D

Spatiotemporally or non-significantly expressed genes (n= 25,979)

GO ID	Description	FDR
GO:0071472	cellular response to salt stress	2E-103
GO:0042631	cellular response to water deprivation	1E-73
GO:0016567	protein ubiquitination	4E-40
GO:0006278	RNA-dependent DNA biosynthetic process	2E-36
GO:0009991	response to extracellular stimulus	1E-33
GO:0070647	protein modification by small protein conjugation or removal	1E-31
GO:0071214	cellular response to abiotic stimulus	1E-29
GO:1900864	mitochondrial RNA modification	1E-25
GO:0071897	DNA biosynthetic process	4E-24
GO:0009451	RNA modification	1E-21
GO:0032953	regulation of (1->3)-beta-D-glucan biosynthetic process	4E-19
GO:0016556	mRNA modification	1E-18
GO:0000959	mitochondrial RNA metabolic process	2E-18
GO:0046620	regulation of organ growth	1E-15
GO:0016553	base conversion or substitution editing	3E-12
GO:0032881	regulation of polysaccharide metabolic process	1E-08
GO:0016554	cytidine to uridine editing	9E-08
GO:0006074	(1->3)-beta-D-glucan metabolic process	8E-07
GO:1901599	(-)-pinoresinol biosynthetic process	5E-06
GO:0080187	floral organ senescence	9E-06
GO:0080112	seed growth	7E-05
GO:0032988	ribonucleoprotein complex disassembly	1E-04
GO:0080113	regulation of seed growth	1E-04
GO:0045697	regulation of synergid differentiation	1E-04
GO:0060688	regulation of morphogenesis of a branching structure	1E-04
GO:0009563	synergid differentiation	1E-04
GO:0051603	proteolysis involved in cellular protein catabolic process	3E-04
GO:0009969	xyloglucan biosynthetic process	3E-04
GO:0071249	cellular response to nitrate	6E-04
GO:0006109	regulation of carbohydrate metabolic process	1E-03

FIGURE 2 | Spatiotemporally or non-significantly expressed genes in hexaploid wheat. **(A)** Distribution of spatiotemporally or non-significantly expressed genes across the 21 wheat chromosomes. **(B)** Proportion of the spatiotemporally expressed genes (RPM ≥ 3 in at least one RNA-seq dataset) and non-significantly expressed genes (RPM < 3 in all of the 727 RNA-seq datasets) in hexaploid wheat. **(C)** t-SNE plot of the spatiotemporally expressed genes. Clusters of genes expressed in roots, stamens, and anthers are circled. **(D)** Enriched GO terms in the biological processes of the spatiotemporally or non-significantly expressed genes in hexaploid wheat.

Wang et al. (2017) found that approximately 60% of wheat genes are expressed from one or two homoeologous loci in reproductive tissues (Wang et al., 2017). Our findings based on more comprehensive transcriptome datasets showed that, compared with previous observations, a smaller number of genes (~40% of genes assigned into each of the homoeologous groups) are expressed from one or two homoeologous loci (**Figure 1**). These observations suggest that approximately 15–20% of wheat genes, including the silenced loci considered in previous studies, may contain homoeologs that can be expressed in specific tissues, at different developmental stages, or under different conditions. Our list of the spatiotemporally expressed and non-significantly expressed genes represent as many as 44% of those genes expressed (RPM ≥ 3) in 1–7 datasets out of the 727 RNA-seq datasets, and suggested that some of these are particularly expressed in specific tissues such as roots, stamens, and anthers (**Figures 2B,C**). Although we used a threshold of RPM ≥ 3 in less than



GO ID	Description	FDR
GO:0009102	biotin biosynthetic process	2E-05
GO:0006768	biotin metabolic process	2E-05
GO:0006476	protein deacetylation	1E-04
GO:0098732	macromolecule deacylation	4E-04
GO:0071370	cellular response to gibberellin stimulus	6E-04
GO:0009250	glucan biosynthetic process	1E-03

FIGURE 3 | Co-expressed while differentially expressed homoeologs in hexaploid wheat. **(A)** Proportion of co-expression patterns of homoeologs in the expressed homoeologous triplets. ABD, homoeologous triplets in which all three homoeologs are grouped in the same co-expression module; AB-D, homoeologous triplets in which A- and B-homoeologs are grouped in the same co-expression module while D-homoeolog is in another co-expression module; AD-B, homoeologous triplets in which A- and D-homoeologs are grouped in the same co-expression module while B-homoeolog is in another co-expression module; BD-A, homoeologous triplets in which B- and D-homoeologs are grouped in the same co-expression module while A-homoeolog is in another co-expression module; A-B-D, homoeologous triplets in which all three homoeologs are assigned to different modules; Not clustered, homoeologous triplets in which two or all three homoeologs are not assigned to a co-expression module. **(B)** Number of co-expressed triplets containing differentially expressed homoeologs across ≥ 50% of the 727 RNA-seq datasets. **(C)** Box plot of the expression levels of the homoeologs in 15 homoeologous triplets showing a consistent trend of homoeolog expression bias ≥ 95% across the 727 RNA-seq datasets. **(D)** Enriched GO terms in the biological processes of genes in the 258 co-expressed triplets containing differentially expressed homoeologs across ≥ 50% of the 727 RNA-seq datasets.

1% (eight datasets) of the 727 RNA-seq datasets to identify spatiotemporally or non-significantly expressed genes, this threshold depends on the proportion of samples from similar tissues in the dataset, which might present genes specifically expressed in unusually sequenced samples. To further explore spatiotemporally expressed genes, transcriptome datasets obtained from anatomically- or seasonally-distinct samples should be analyzed using emerged technologies such as laser-capture microdissection RNA-seq (LCM RNA-seq) (Zhan et al., 2015) and field transcriptome sequencing (Plessis et al., 2015). These findings may suggest that such genes expressed only from one or two homoeoalleles undergo transcriptional silencing, probably through differentiation of expression patterns and specialization of spatial expression. Consequently, such duplicated genes might be non-functionalized through promoter malfunctions or repression of other transcriptional machineries as a process of functional diploidization (Levy and Feldman, 2002; Rajkov et al., 2014).

Our gene co-expression network analysis enabled us to identify homoeologous triplets containing homoeologs that are co-expressed while differentially expressed (2.5% of the 10,415 expressed homoeologous triplets), as well as differentiated homoeologs that are classified into co-expression modules that differ from the other two homoeologs (27% of the 10,415 expressed homoeologous triplets) (**Figures 3A,B**). The results of our comprehensive analysis provide evidence that may suggest that most of the differential expression observed between homoeologs represents an alteration of expression patterns in hexaploid wheat. The results of our co-expressed gene network analysis enable us to identify transcriptional modules that contain abundant differentiated homoeologs involved in several particular biological processes, which might have evolved such biological functions in hexaploid wheat through its allopolyploidization (Chen et al., 2007; Feldman and Levy, 2012). Multiple studies have provided evidence to suggest that homoeolog subfunctionalization may be related to enhanced

A

Module 1	Module 2	Module 3	Module 4	Module 5	Module 6	Module 7	Module 8
14,940 (585, 4%)	2,733 (205, 8%)	2,070 (236, 11%)	2,047 (238, 12%)	1,808 (166, 9%)	1,548 (176, 11%)	1,364 (273, 20%)	838 (238, 28%)

Module 9	Module 10	Module 11	Module 12	Module 13	Module 14	Module 15	Module 16
815 (146, 18%)	471 (98, 21%)	351 (38, 11%)	308 (18, 6%)	262 (43, 16%)	169 (24, 14%)	166 (37, 22%)	122 (12, 10%)

Module 17	Module 18	Module 19	Module 20	Module 21	Module 22	Module 0	Total
122 (12, 10%)	82 (18, 22%)	80 (15, 19%)	76 (19, 25%)	45 (10, 22%)	41 (5, 12%)	787 (219, 28%)	31,245 (2,831, 9%)

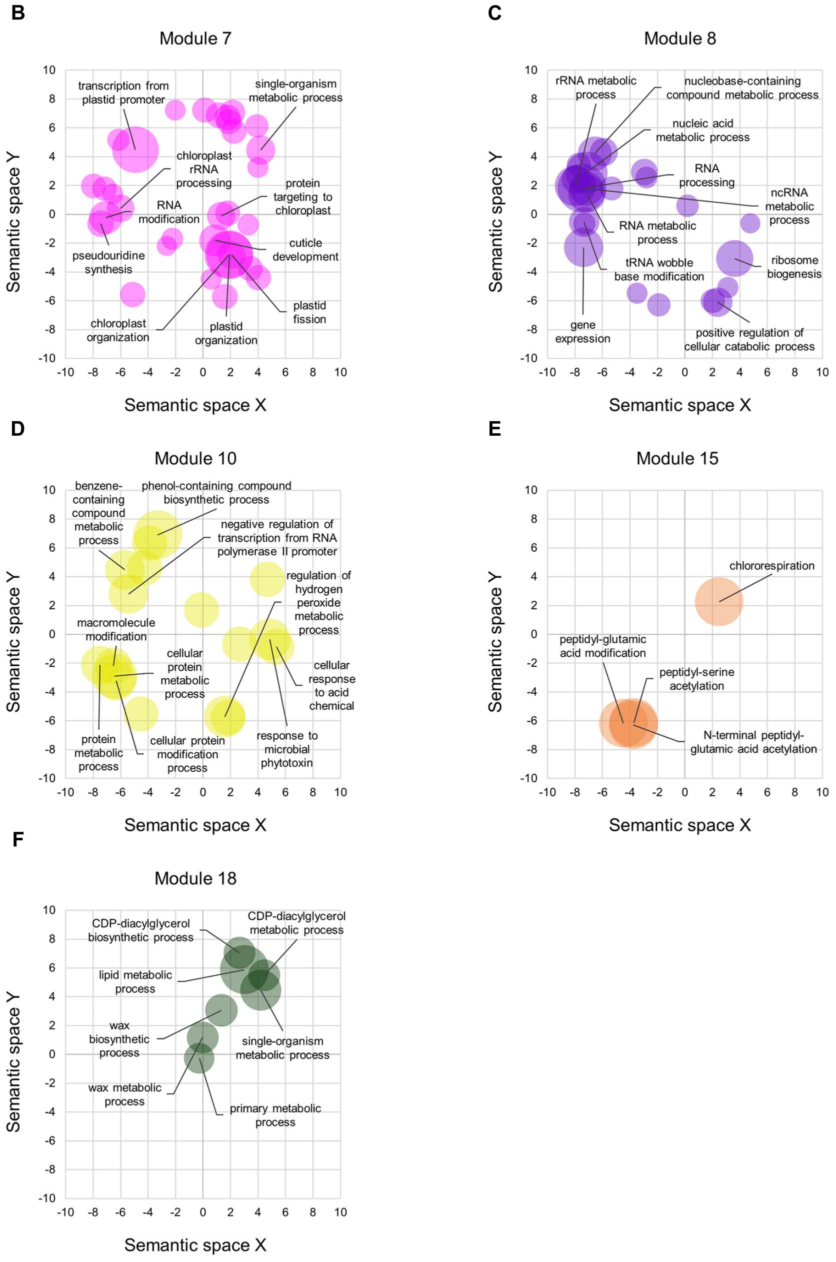


FIGURE 4 | Co-expression modules containing the differentiated homoeologs in hexaploid wheat. **(A)** Number of homoeologs and differentiated homoeologs (numbers with their percentage in brackets) in each of the co-expression modules. The Module 0 represents genes that are not clustered in a co-expression module. Percentages represent proportions of the differentiated homoeologs in each of the co-expression modules. **(B–E)** Enriched GO terms in the biological processes of genes in the co-expression module 7 **(B)**, 8 **(C)**, 10 **(D)**, 15 **(E)**, and 18 **(F)** projected to a 2D semantic space. Circle size represents the –log10 of FDR values calculated using REVIGO analysis. The top ten enriched GO terms are labeled in the plots.

adaptability to adverse environmental conditions in various allopolyploid species, such as tetraploid cotton, tetraploid coffee, and hexaploid wheat (Liu and Adams, 2007; Hu et al., 2011; de Carvalho et al., 2014; Liu et al., 2015). Consequently, our results suggest that along with other genes, such differentiated homoeologs may have innovated transcriptional networks, which may have contributed to adaptation to environmental change as well as to enhanced productivity during the evolution of hexaploid wheat.

The large number of RNA-seq datasets analyzed in the current study allowed integrating the transcriptional properties of each homoeologous triplet into a dataset (**Supplementary Table S3**), thereby providing a useful information resource for understanding the evolution and function of duplicated genes in hexaploid wheat. Moreover, our analyses using the datasets enabled us to demonstrate the presence of co-expression modules containing a high proportion of differentiated homoeologs in hexaploid wheat, which in turn allowed us to dissect its complex transcriptome derived from duplicated genomes. The considerable recent advances in whole-genome assembly in Triticeae species, including hexaploid wheat and its ancestors (Ling et al., 2013; Mochida and Shinozaki, 2013; International Wheat Genome Sequencing Consortium, 2014; Luo et al., 2017), provide us with an opportunity to further explore sub-/neofunctionalized homoeologs and elucidate the diploidization process that occurred during the evolution of hexaploid wheat after allopolyploidization. Such analysis will enable us to identify genes and transcriptional modules that may be associated with adaptive traits in hexaploid wheat. Such genes and transcriptional modules might also prove useful in enhancing the adaptation of staple crops to counter the potentially adverse impacts of global climate changes and improve their productivity.

AUTHOR CONTRIBUTIONS

KT and KM designed the work. KT and KI performed the bioinformatics analysis. KT and KM wrote the manuscript.

FUNDING

This work was partially supported by grants-in-aid for Young Scientists (A) (Grant No. 26712003 to KM) from the Japan Society for the Promotion of Science (JSPS) and by funds to KM from the CREST (JPMJCR16O4) of the Japan Science and Technology Agency (JST).

ACKNOWLEDGMENTS

The authors thank Shun Takaya for his help in data analysis. This work was supported by RIKEN Junior Research Associate Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01163/full#supplementary-material

REFERENCES

Adams, K. L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered.* 98, 136–141. doi: 10.1093/jhered/esl061

Adams, K. L., and Wendel, J. F. (2005). Novel patterns of gene expression in polyploid plants. *Trends Genet.* 21, 539–543. doi: 10.1016/j.tig.2005.07.009

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691. doi: 10.1105/tpc.021410

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

Chaudhary, B., Flagel, L., Stupar, R. M., Udall, J. A., Verma, N., Springer, N. M., et al. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). *Genetics* 182, 503–517. doi: 10.1534/genetics.109.102608

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* 15, 57–71. doi: 10.1016/j.tplants.2009.12.003

Chen, Z. J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* 28, 240–252. doi: 10.1002/bies.20374

Chen, Z. J., Ha, M., and Soltis, D. (2007). Polyploidy: genome obesity and its consequences. *New Phytol.* 174, 717–720. doi: 10.1111/j.1469-8137.2007.02084.x

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Combes, M. C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New Phytol.* 200, 251–260. doi: 10.1111/nph.12371

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, 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

Cusack, B. P., and Wolfe, K. H. (2007). When gene marriages don't work out: divorce by subfunctionalization. *Trends Genet.* 23, 270–272. doi: 10.1016/j.tig.2007.03.010

de Carvalho, K., Petkowicz, C. L., Nagashima, G. T., Bespalhok Filho, J. C., Vieira, L. G., Pereira, L. F., et al. (2014). Homeologous genes involved in mannitol synthesis reveal unequal contributions in response to abiotic stress in *Coffea arabica*. *Mol. Genet. Genomics* 289, 951–963. doi: 10.1007/s00438-014-0864-y

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., et al. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* 42, 443–461. doi: 10.1146/annurev.genet.42.110807.091524

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

Feldman, M., and Levy, A. A. (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* 192, 763–774. doi: 10.1534/genetics.112.146316

Hu, Z., Yu, Y., Wang, R., Yao, Y., Peng, H., Ni, Z., et al. (2011). Expression divergence of TaMBD2 homoeologous genes encoding methyl CpG-binding domain proteins in wheat (*Triticum aestivum* L.). *Gene* 471, 13–18. doi: 10.1016/j.gene.2010.10.001

Hughes, T. E., Langdale, J. A., and Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* 24, 1348–1355. doi: 10.1101/gr.172684.114

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788. doi: 10.1126/science.1251788

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326. doi: 10.1101/gr.101386.109

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

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 Genomics* 15:276. doi: 10.1186/1471-2164-15-276

Leitch, A. R., and Leitch, I. J. (2008). Perspective – Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–483. doi: 10.1126/science.1153585

Levy, A. A., and Feldman, M. (2002). The impact of polyploidy on grass genome evolution. *Plant Physiol.* 130, 1587–1593. doi: 10.1104/pp.015727

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

Ling, H. Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., et al. (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* 496, 87–90. doi: 10.1038/nature11997

Liu, Z., and Adams, K. L. (2007). Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. *Curr. Biol.* 17, 1669–1674. doi: 10.1016/j.cub.2007.08.030

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

Luo, M. C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., et al. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* 551, 498–502. doi: 10.1038/nature24486

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity (Edinb)* 110, 99–104. doi: 10.1038/hdy.2012.79

Matsuoka, Y. (2011). Evolution of polyploid triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol.* 52, 750–764. doi: 10.1093/pcp/pcr018

McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, W20–W25. doi: 10.1093/nar/gkh435

Mochida, K., and Shinozaki, K. (2013). Unlocking triticeae genomics to sustainably feed the future. *Plant Cell Physiol.* 54, 1931–1950. doi: 10.1093/pcp/pct163

Mochida, K., Yamazaki, Y., and Ogihara, Y. (2003). Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genomics* 270, 371–377. doi: 10.1007/s00438-003-0939-7

Moore, R. C., and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* 8, 122–128. doi: 10.1016/j.pbi.2004.12.001

Naseeb, S., Ames, R. M., Delneri, D., and Lovell, S. C. (2017). Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc. Biol. Sci.* 284:20171393. doi: 10.1098/rspb.2017.1393

Osabe, K., Kawanabe, T., Sasaki, T., Ishikawa, R., Okazaki, K., Dennis, E. S., et al. (2012). Multiple mechanisms and challenges for the application of allopolyploidy in plants. *Int. J. Mol. Sci.* 13, 8696–8721. doi: 10.3390/ijms13078696

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., et al. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* 345:1250091. doi: 10.1126/science.1250091

Plessis, A., Hafemeister, C., Wilkins, O., Gonzaga, Z. J., Meyer, R. S., Pires, I., et al. (2015). Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. *Elife* 4:e08411. doi: 10.7554/eLife.08411

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

Rajkov, J., Shao, Z., and Berrebi, P. (2014). Evolution of polyploidy and functional diploidization in sturgeons: microsatellite analysis in 10 sturgeon species. *J. Hered.* 105, 521–531. doi: 10.1093/jhered/esu027

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Renny-Byfield, S., and Wendel, J. F. (2014). Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* 101, 1711–1725. doi: 10.3732/ajb.1400119

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

Solhaug, E. M., Ihinger, J., Jost, M., Gamboa, V., Marchant, B., Bradford, D., et al. (2016). Environmental regulation of heterosis in the allopolyploid *Arabidopsis suecica*. *Plant Physiol.* 170, 2251–2263. doi: 10.1104/pp.16.00052

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Takahagi, K., Inoue, K., Shimizu, M., Uehara-Yamaguchi, Y., Onda, Y., and Mochida, K. (2018). Homoeolog-specific activation of genes for heat acclimation in the allopolyploid grass *Brachypodium hybridum*. *Gigascience* 7:giy020. doi: 10.1093/gigascience/giy020

Tanaka, M., Tanaka, H., Shitsukawa, N., Kitagawa, S., Takumi, S., and Murai, K. (2016). Homoeologous copy-specific expression patterns of MADS-box genes for floral formation in allopolyploid wheat. *Genes Genet. Syst.* 90, 217–229. doi: 10.1266/ggs.15-00029

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Wang, Y., Yu, H., Tian, C., Sajjad, M., Gao, C., Tong, Y., et al. (2017). Transcriptome association identifies regulators of wheat spike architecture. *Plant Physiol.* 175, 746–757. doi: 10.1104/pp.17.00694

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Mol. Biol.* 42, 225–249. doi: 10.1023/A:1006392424384

Whittle, C. A., and Krochko, J. E. (2009). Transcript profiling provides evidence of functional divergence and expression networks among ribosomal protein gene paralogs in *Brassica napus*. *Plant Cell* 21, 2203–2219. doi: 10.1105/tpc.109.068411

Yang, C., Zhao, L., Zhang, H., Yang, Z., Wang, H., Wen, S., et al. (2014). Evolution of physiological responses to salt stress in hexaploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11882–11887. doi: 10.1073/pnas.1412839111

Zhan, J., Thakare, D., Ma, C., Lloyd, A., Nixon, N. M., Arakaki, A. M., et al. (2015). RNA sequencing of laser-capture microdissected compartments of the maize kernel identifies regulatory modules associated with endosperm cell differentiation. *Plant Cell* 27, 513–531. doi: 10.1105/tpc.114.135657

Conflict of Interest Statement: 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 07 August 2018
doi: 10.3389/fevo.2018.00114

Adding Complexity to Complexity: Gene Family Evolution in Polyploids

Barbara K. Mable[1]*, *Anne K. Brysting*[2], *Marte H. Jørgensen*[2], *Anna K. Z. Carbonell*[1,3], *Christiane Kiefer*[4], *Paola Ruiz-Duarte*[4], *Karin Lagesen*[2,5] and *Marcus A. Koch*[4]

[1] Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, United Kingdom, [2] Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway, [3] Department of Biological and Environmental Sciences, University of Stirling, Stirling, United Kingdom, [4] Centre for Organismal Studies Heidelberg, Department of Biodiversity and Plant Systematics, Botanic Garden and Herbarium Heidelberg, University of Heidelberg, Heidelberg, Germany, [5] Department of Bioinformatics, University of Oslo, Oslo, Norway

Comparative genomics of non-model organisms has resurrected whole genome duplication (WGD) from being viewed as a somewhat obscure process that happens in plants to a primary driver of eukaryotic diversification. The shadow of past ploidy increases has left a strong signature of duplicated genes organized into gene families, even in small genomes that have undergone effectively complete rediploidization. Nevertheless, despite continually advancing technologies and bioinformatics pipelines, resolving the fate of duplicate genes remains a substantial challenge. For example, many important recognition processes are driven not only by allelic expansion through retention of duplicates but also by diversification and copy number variation. This creates technical difficulties with assembly to reference genomes and accurate interpretation of homology. Thus, relatively little is known about the impacts of recent polyploidization and hybridization on the evolution of gene families under selective forces that maintain diversity, such as balancing selection. Here we use a complex of species and ploidy levels in the genus *Arabidopsis* (*A. lyrata* and *A. arenosa*) as a model to investigate the evolutionary dynamics of a large and complicated gene family known to be under strong balancing selection: the receptor-like kinases, which include the female component of genetically controlled self-incompatibility. Specifically, we question: (1) How does diversity of *S*-receptor kinase (*SRK*) alleles in tetraploids compare to that in their close diploid relatives? (2) Is there increased trans-specific polymorphism (i.e., sharing of alleles that transcend speciation, characteristic of balancing selection) in tetraploids compared to diploids due to the higher number of copies they carry? (3) Do these highly variable loci show evidence of introgression among extant species/ploidy levels within or outside known zones of hybridization? (4) Is there evidence for copy number variation among paralogs? We use this example to highlight specific issues to consider when interpreting gene family evolution, particularly in relation to polyploids but also more generally in diploids. We conclude with recommendations for strategies to address the challenges of resolving such complex loci in the future, using advances in deep sequencing approaches.

Keywords: polyploidy, gene family evolution, self-incompatibility, copy number variation, trans-specific polymorphism, balancing selection, introgression

OPEN ACCESS

Edited by:
Richard John Abbott,
University of St Andrews,
United Kingdom

Reviewed by:
Céline Poux,
Université de Lille, France
Baocheng Guo,
Institute of Zoology (CAS), China

****Correspondence:***
Barbara K. Mable
barbara.mable@glasgow.ac.uk

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

Received: *29 March 2018*
Accepted: *17 July 2018*
Published: *07 August 2018*

Citation:
Mable BK, Brysting AK,
Jørgensen MH, Carbonell AKZ,
Kiefer C, Ruiz-Duarte P, Lagesen K
and Koch MA (2018) Adding
Complexity to Complexity: Gene
Family Evolution in Polyploids.
Front. Ecol. Evol. 6:114.
doi: 10.3389/fevo.2018.00114

INTRODUCTION

Background and Aims

The sequencing of the human genome in 2001 (Lander et al., 2001) promised to revolutionize modern medicine and lead to a new era in understanding the complexity of genetic control of complex phenotypes. While this has certainly been true, it is really the comparative genomics of non-model organisms that has led to a complete revolution in understanding (e.g., Seeb et al., 2011; da Fonseca et al., 2016). One unexpected finding was that whole genome duplication (WGD) has been an important process contributing to the genomic history of all eukaryotes, including those with relatively small genomes, such as the yeast *Saccharomyces cerevisiae* (Wolfe and Shields, 1997) and the model plant *Arabidopsis thaliana* (Blanc and Wolfe, 2004). Although Susumu Ohno in the late 1960s had emphasized the central role of gene duplication in the evolutionary history of vertebrates (Ohno, 1970), it wasn't until after his death in 2000 that comparative genomic studies confirmed that fish had undergone multiple rounds of WGD (e.g., Meyer and Van de Peer, 2005), as he had predicted. He also had predicted that effective rediploidization following duplication was inevitable but that some duplicates would be retained to perform new or specialized functions, leaving a footprint of past duplications and organization of genes into gene families. His ideas about the fates of duplicate genes to include specialization of function (now known as "subfunctionalization"; Force et al., 1999) also have been resurrected and form the basis for understanding the history of complex genomes such as salmonids, which underwent an independent WGD after the last teleost specific duplication (Hermansen et al., 2016; Lien et al., 2016). Comparative studies of vertebrates have thus been critical for establishing polyploidization as a creative evolutionary force shaping the genomes of all eukaryotes (Van de Peer et al., 2017), as had long been recognized for plants (e.g., Soltis et al., 1992; Adams, 2007).

Nevertheless, despite recognition that duplicated genes are critical for understanding genome structure and function (Van de Peer et al., 2017), the practicalities of assembling duplicates in genomic resequencing studies, resolving orthology, and interpreting their potentially redundant effects on phenotypes remains a substantial challenge (da Fonseca et al., 2016). Retention of duplicate genes following genomic or tandem duplication is non-random (Adams, 2007) and is both constrained and promoted by achieving appropriate levels of expression (e.g., Gout and Lynch, 2015; Mattenberger et al., 2017; Rodrigo and Fares, 2018). The "gene balance" hypothesis, for example, predicts that loci involved in regulating levels of expression of integrated genetic pathways (such as transcription factors or members of signal transduction pathways) should show increased retention of duplicates to maintain coordinated function (Birchler and Veitia, 2010). Genes for which high expression is advantageous might be expected to retain expression in duplicated copies whereas divergence in patterns of expression could be advantageous for others. Genes that are retained in duplicate through one round of WGD also have been found to be preserved through later rounds (Seoighe and Gehring, 2004). Thus, not considering the role of gene copies retained in duplicate could alter interpretation of regulatory processes associated with adaptation.

One type of adaptive process often associated with large and complex gene families is recognition of self vs. non-self, where high polymorphism is favored by continually changing selection pressures, and retention of duplicate copies could be beneficial for increasing allelic repertoire. For example, the "big bang" theory of the emergence of the adaptive immune systems in vertebrates invokes multiple rounds of WGD as the major source of this potential (Flajnik and Kasahara, 2010). Similarly, investigation of the genomic repertoire of pathogen-associated genes (R genes) in several crop plants through targeted sequence capture (Jupe et al., 2012, 2013; Giolai et al., 2016; Van Weymers et al., 2016) has revealed much more extensive gene families than was previously predicted based on whole genome resequencing studies. R genes have also been demonstrated to show signatures of adaptive introgression between closely related species of *Arabidopsis*, with extensive trans-specific sharing of alleles across species (Bechsgaard et al., 2017). An added complication for these types of gene families is that copy number can be variable even among individuals within a species (e.g., Mable et al., 2015), meaning that genome references will not always include the full complement of copies. Copy number variation has been linked to disease severity in humans (Beckmann et al., 2007; Wheeler et al., 2008) and adaptive processes in other organisms (Saintenac et al., 2011; Zmienko et al., 2014; Duvaux et al., 2015; Hull et al., 2017) but methods that can reliably distinguish between lack of coverage and variation in presence of a particular gene copy are required to fully evaluate the evolutionary significance of presence/absence polymorphisms following gene duplication.

The high polymorphism expected for recognition genes means that they are prime candidates to be "lost" in genomic resequencing studies, even in diploids. For example, genes controlling sporophytically controlled self-incompatibility (SI) in plants have been found to be missing from resequencing assemblies because they are too divergent from the reference genome and so trawling in the unassembled reads is necessary to characterize these highly polymorphic genes (Mable et al., 2017). Both male and female components are members of large gene families that show extensive trans-specific polymorphism, with highly similar alleles shared across species and even genera but high divergence between functional specificities (Schierup et al., 1998; Paetsch et al., 2006; Castric and Vekemans, 2007; Busch et al., 2008; Guo et al., 2011; Tedder et al., 2011; Leducq et al., 2014). The gene controlling female specificity (*S*-receptor kinase, *SRK*) is part of a large family of receptor kinases, which evolved through a complex history of gene duplication and loss, followed by gene fission and fusion (Xing et al., 2013). Gene conversion between *SRK* and other members of the gene family is also thought to have contributed to expansion of functional allelic diversity (Prigoda et al., 2005; Guo et al., 2011). This creates additional challenges with interpreting which variants are parts of the functional locus regulating the SI response and which are functionally unlinked but show high sequence similarity. For sporophytic SI, the phenotype of the pollen is determined by the genotype of the diploid (or tetraploid) parent, so there can be dominance in both pollen and stigma.

Dominance is known to be complex, with non-linear interactions that can differ between pollen and stigma (Lewis, 1947; Stevens and Kay, 1989; Hatakeyama et al., 1998; Shiba et al., 2002; Mable et al., 2003; Llaurens et al., 2009; Schoen and Busch, 2009). Trans-specific polymorphism (i.e., sharing of alleles that transcends speciation) of *SRK* alleles has been well established for diploids (Charlesworth et al., 2006; Boggs et al., 2009; Castric et al., 2010), and is thought to be a key indicator of the action of balancing selection (Takahata, 1990). However, the strength of balancing selection on tetraploids has not been assessed specifically. Since tetraploid individuals can carry up to four different *SRK* alleles, there is potential for increased sharing across species, at least of recessive alleles. They can also carry multiple copies of recessive alleles (Mable et al., 2004), which could result in the maintenance of more variants within specificities than for diploids. While previous work has demonstrated that linkage and dominance works similarly in tetraploid compared to diploid *Arabidopsis lyrata* (Mable et al., 2004), the evolutionary dynamics of *S*-alleles in tetraploids has not been studied.

In addition, interpreting the fate of duplicate genes in polyploids is complicated by the fact that hybridization is often associated with WGD and so it can be difficult to disentangle the effects of combining and duplicating genomes on patterns of duplicate gene expression or dynamics of gene families (e.g., Evans, 2007; Guggisberg et al., 2009; Mable, 2013). Fortunately, rapid advances in sequencing technology and bioinformatic processing mean that the toolbox available to resolve such challenges continues to improve. Targeted sequence capture, for example, has been used effectively to investigate genomic changes in polyploids (Salmon et al., 2012; Gardiner et al., 2016; Krasileva et al., 2017). However, even with these advances in technology there are important issues to consider when resolving and interpreting evolutionary dynamics of gene families, particularly for systems in which recent polyploidization and hybridization could complicate accurate assembly into orthologs and subsequent genotyping within and between copies.

The purpose of this paper is to discuss these issues in the context of understanding the evolutionary dynamics of the *SRK* gene family in a species complex (*A. lyrata* and *A. arenosa*) that includes both diploids and tetraploids, with tetraploids showing extensive introgression in a hybrid zone in central Europe (Schmickl et al., 2010; Jørgensen et al., 2011; Schmickl and Koch, 2011; Hohmann et al., 2014; Muir et al., 2015; Novikova et al., 2016; Hohmann and Koch, 2017). In *A. arenosa,* tetraploids have been predicted to have arisen through autopolyploidisation (Arnold et al., 2015); secondary contact with *A. lyrata* during interglacial and postglacial range contractions and expansions has subsequently led to introgression between tetraploids in the two species. Our intent was to use investigation of *S*-receptor kinase evolution in this species complex as a model for understanding how balancing selection operates in polyploid genomes and to determine whether these highly polymorphic gene families could be useful indicators of hybridization and introgression. Specifically, our objectives were to question: (1) How does diversity of *SRK*-related alleles in tetraploids compare to that in their close diploid relatives? (2) Is there increased trans-specific polymorphism of *SRK* alleles in tetraploids compared to diploids because of the increased number of copies they can carry? (3) Do these highly variable loci show evidence of introgression among extant species/ploidy levels within or outside known zones of hybridization? (4) Is there evidence for copy number variation among paralogs?

We use these questions to highlight the challenges for interpreting gene family evolution, particularly in polyploids, but also relevant to diploids. We conclude with recommendations for how some of these challenges might be overcome using deep sequencing approaches. We reiterate the recommendation from others (Salmon et al., 2012; Jupe et al., 2013; Gardiner et al., 2016; Van Weymers et al., 2016; Krasileva et al., 2017) that non-amplicon based targeted sequence capture (e.g., whole genome exon capture or targeting of particular gene families) is the most promising method for tackling the full complexity of gene family evolution in complex genomes but suggest cautionary strategies that should be considered when interpreting evolutionary patterns.

Notes on Terminology and Known Challenges Associated With the *SRK* Gene Family

A complication with understanding the evolution of complex gene families is distinguishing what is meant by an "allele." For *SRK,* there can be sequence variation within "specificities," which are *SRK* types that confer a specific SI phenotype (i.e., a protein expressed on the surface of the stigma that is recognized as self by the comparable protein expressed on the surface of the pollen grain). These specificities (which we will refer to as "alleles") can be as divergent from one another as they are from other genes (which we will refer to as "loci") in the same gene family. Moreover, phylogenetic clustering alone is not sufficient to predict which sequence variants represent *SRK* alleles because gene conversion with unlinked loci has resulted in higher similarity between paralogs than among *SRK* alleles (Prigoda et al., 2005). Diploid individuals should contain only two functional *SRK* alleles but could contain varying numbers of loci in the gene family that are not linked to the SI phenotype; since tetraploids can contain multiple copies of the same allele without altering the specificity or dominance (Mable et al., 2004), the number of *SRK* alleles expected in a polyploid cannot be predicted. Thus, assigning "sequence variants" to gene family loci or *SRK* alleles is even more complicated in polyploids than for diploids. *SRK* alleles have been grouped into four different dominance classes (A1, A2, A3, B; Prigoda et al., 2005). Polymorphisms within specificities/alleles (which we will refer to as "haplotypes") are more apparent for recessive than dominant alleles because the former are expected to occur at higher frequency and show more sharing between populations (Bechsgaard et al., 2006; Castric and Vekemans, 2007; Castric et al., 2008, 2010; Stoeckel et al., 2008; Llaurens et al., 2009; Goubet et al., 2012). There is a single most recessive allele (S1, Class A1; Prigoda et al., 2005) that is found globally and in multiple species in the genus *Arabidopsis* (Mable et al., 2003; Dart et al., 2004; Prigoda et al., 2005; Mable and Adam, 2007; Castric et al., 2010; Foxe et al., 2010). Alleles in Class B are recessive to all other classes except S1 but are more similar to unlinked loci (*Aly*13-2 and *Aly*13-7) than to the other classes (Prigoda et al., 2005) and show more intra-allele polymorphisms than dominant alleles (Classes A2 and A3; Prigoda et al., 2005; Castric et al., 2010). The high trans-specific polymorphism also means that naming of alleles can be confusing because a variant found in certain species is often provided a specific number before discovering that it potentially represents the same specificity as an already named allele in another species (Castric et al., 2010). Thus, alleles are named with the species in which they were originally described as a prefix (e.g., *Aly* refers to *A. lyrata*, *Aha* refers to *A. halleri*, *Ath* refers to *A. thaliana*, *Aar* refers to *A. arenosa*). Finally, since the SI phenotype is determined by a combination of variants at the female *SRK* and male *SCR* genes, phenotypic specificities are labeled only "S#" (e.g., S1) for segregation analyses.

From our previous studies on the evolutionary dynamics of *SRK* alleles in diploids, we have already described challenges in generating robust data for interpreting these complex gene families in diploids, relevant for the sequencing strategies we apply here: (1) Primers designed to be general enough to recognize all *SRK* alleles also amplify the rest of the gene family, so a major challenge is assigning sequence variants to loci (Schierup et al., 2001; Charlesworth et al., 2003b; Mable et al., 2003, 2017; Mable and Adam, 2007). (2) This is complicated by the fact that, due to the extensive polymorphism in *SRK* and evidence that gene conversion has contributed to allelic repertoire, paralogs that are not linked to the SI phenotype can be more similar to "real" alleles than "real" alleles are to one another, so similarity can't always be used to assign functionality (Schierup et al., 2001; Mable et al., 2003; Prigoda et al., 2005). (3) Amplicon-based approaches are inherently at risk of generating PCR recombinants between copies, making it difficult to distinguish errors from actual recombination, introgression in hybrids, or gene conversion between sequences. (4) It is also difficult to distinguish presence/absence of paralogs from amplification biases during PCR (Mable et al., 2017). (5) There is extensive length heterogeneity within and between members of the gene family, so it can be difficult to establish the positional homology necessary to interpret patterns of selection (Charlesworth et al., 2003a). (6) The highly polymorphic nature of *SRK* alleles means that they are sometimes too divergent from the reference genome to be assembled using standard filtering strategies; this means that these types of alleles might frequently be found in the unassembled reads for resequencing projects (Mable et al., 2017).

MATERIALS AND METHODS

Sampling and Overview of Methods

Samples were obtained from both diploid and tetraploid populations of *A. lyrata* and *A. arenosa* sampled from Central Europe (**Table 1**). Although current systematics suggests separation of diploid and tetraploid *A. arenosa* into distinct species taxonomically (Koch et al., 2008), for simplicity, we will refer to both as *A. arenosa* here. We sampled individuals from 3-5 populations of each "type": A2x refers to diploid *A. arenosa*, A4x to tetraploid *A. arenosa*, L2x to diploid *A. lyrata* and L4x to tetraploid *A. lyrata*. Tetraploid populations occurring in a hybrid zone between the two species (Schmickl, 2009; Schmickl and Koch, 2011; Hohmann et al., 2014; Muir et al., 2015; Novikova et al., 2016) were included to test for patterns of introgression. Diploids have not been found to hybridize (Jørgensen et al., 2011) and so were considered "pure" populations. To test patterns of linkage of sequence variants with the SI phenotype, we also included 104 individuals from crosses between *A. lyrata* tetraploid parents whose genotypes had been partially resolved by cloning and Sanger sequencing; we performed di-allele crosses within these families to establish SI phenotypes that could be compared to the 454 genotypes.

We used a combination of approaches to address the main research questions: (1) 454 pyrosequencing using degenerate primers (**Supplementary Table 1**) targeting the *SRK* gene family (Jørgensen et al., 2012) to characterize diversity and patterns of allele sharing in diploids and polyploids; (2) direct Sanger sequencing to investigate signatures of introgression in shared haplotypes and for segregation analyses to test linkage to the SI phenotype; (3) cloning and Sanger sequencing using degenerate primers (**Supplementary Table 1**) to obtain longer products than possible with 454 pyrosequencing to further characterize potentially new alleles; and (4) using data from a recent genomic resequencing study (Novikova et al., 2016) to search for the *SRK* gene family using novel assembly approaches, to test whether copy number variation and patterns of introgression can be mined using existing genomic data. We focused on variation in exon 1 (the *S*-domain) because it contains the sites used for recognition of self vs. non-self (Schierup et al., 2001; Charlesworth et al., 2003a). However, we also used the genome mining approach to determine whether we could pull out full-length sequences that include the functional kinase domain (exons 3-7).

While 454 pyrosequencing has largely been replaced by methods demonstrated to show higher accuracy such as Illumina (Schirmer et al., 2015, 2016; D'Amore et al., 2016), we use results from this study as a platform to highlight considerations for working with gene families that should apply across methods. We thus haven't focused on attempting to resolve 454 specific problems but instead on general issues with clustering and assigning sequence variants to loci and designating allelic specificities for interpretation of gene family evolution. We include these as "challenges" in relation to the methods used to address each objective.

Detailed Methodology

Clustering and *SRK* Genotyping Strategies

To increase the probability of amplifying all variants of *SRK* present in the populations sampled, we used 454 pyrosequencing of pooled amplicons from four sets of degenerate primers but sharing a common reverse sequence, *SLGR* (**Supplementary Table 1**; Schierup et al., 2001). Detailed

TABLE 1 | Populations sampled, indicating the code (Pop Code) used to identify populations in our study, the population identifier (identity) from Schmickl (2009), site description, ploidy, species, country of origin, GPS coordinates (latitude and longitude), and whether the population is in the known hybrid zone in Austria, as well as sample sizes for the 454 pyrosequencing (N 454) and targeted amplicon sequencing of *SRK*01 (N *SRK*01).

Pop Code	Site Description	Identity	Species	Ploidy	Country	Latitude	Longitude	Hybrid zone	N 454	N *SRK*01
A2_SVK1	Vsoky Tatry	131R	*A. arenosa*	Diploid	Slovakia	49.2325	20.1980	No	28	28
A2_SVK2	Velkra Fatra	915141	*A. arenosa*	Diploid	Slovakia	48.8242	19.0233	No	25	0[b]
A2_SVK3	Nizke Tatry	915140	*A. arenosa*	Diploid	Slovakia	48.8843	20.2485	No	25	16
A4_AUT1	Kernhof	915142	*A. arenosa*	Tetraploid	Austria	47.8162	15.5435	Yes	25	18
A4_AUT2	Achleichten, Wachau	123R	*A. arenosa*	Tetraploid	Austria	48.4064	15.4728	Yes	26	16
A4_AUT3	Kamptal	3R	*A. arenosa*	Tetraploid	Austria	48.5306	15.6915	Yes	15	13
A4_AUT4	Scheibenbach, Wachau	89R	*A. arenosa*	Tetraploid	Austria	48.4137	15.5200	Yes	14	11
A4_GER	Wental	20R	*A. arenosa*	Tetraploid	Germany	48.7335	10.0193	No	7	0[b]
L2_AUT1	Pernitz-Pottenstein	112R	*A.* lyrata	Diploid	Austria	47.9275	15.9861	No	25	23
L2_AUT2	Vöslauer Hütte	96R	*A. lyrata*	Diploid	Austria	47.9803	16.1650	No	25	9
L2_CZE	Oslavany, Brno	915143	*A. lyrata*	Diploid	Czech R.	49.1219	16.3244	No	9	8
L2_GER	Veldensteiner Forst	915145	*A. lyrata*	Diploid	Germany	49.6453	11.4508	No	17	17
L4_AUT1	Dürnstein, Wachau	13R	*A. lyrata*	Tetraploid	Austria	48.3970	15.5345	Yes	25	7
L4_AUT2	Mödling	915144	*A. lyrata*	Tetraploid	Austria	48.0768	16.2698	Yes	25	18
L4_AUT3	Bachamsdorf, Wachau	50R	*A. lyrata*	Tetraploid	Austria	48.3722	15.4542	Yes	25	22
L4_AUT4	Lilienfeld	116R	*A. lyrata*	Tetraploid	Austria	47.9981	15.5736	No	21	10
L4_AUT5	Rauheneck Ruin	na[a]	*A. lyrata*	Tetraploid	Austria	48.0021	16.2309	No	19	19
L4_AUT2 x L4_AUT5	Crosses		*A. lyrata*	Tetraploid	Austria			No	104	99
Total									460	334

Crosses performed between individuals sampled from Mödling and Rauheneck Ruin near Baden were used to test segregation of genotypes resolved using 454 and SI phenotypes.

[a] *Not included in Schmickl (2009) but collected from Rauheneck Ruin, near Baden.*

[b] *Insufficient DNA remained after the 454 sequencing to screen for AlySRK01.*

methods for the 454 analyses are described in Jørgensen et al. (2012), including estimation of error rates and the use of segregation within known families to test the reliability of genotyping. The initial paper described the strategies used for clustering reads into contigs and filtering to reduce errors. We recommended that optimal clustering was obtained with a 90% sequence similarity criterion and excluding sequences present at a frequency of <7% of the total reads for an individual; these conclusions were based on a subset of the original data that included repeated runs involving the same individuals. We also recommended that clustering should be conducted after reads were trimmed to 200 bp from the "common primer" end (*SLGR* in this case).

Although the crosses between tetraploid *A. lyrata* individuals confirmed presence of the expected *SRK* alleles known to be present in the parents, they also indicated some inaccuracy in allele calls in relation to barcodes; a number of alleles that were not in the parents were assigned to individuals from the crosses, sometimes at high read numbers (see Jørgensen et al., 2012). We concluded that this was due to tag switching between barcodes, as had been suggested from other studies (van Orsouw et al., 2007; Carlsen et al., 2012). Blank lanes (negative controls) also sometimes contained sequences matching known *SRK* alleles, again often at high read numbers. We thus modified our filtering and clustering strategies in the analysis of the full dataset.

Reads were initially assembled into contigs based on clustering to sequences from a reference database of known *SRK* alleles and known members of the gene family that have been characterized in other studies and from our unpublished data from Sanger sequencing. A second iteration then used newly sequenced reads as seeds for clustering, in order to identify putatively new alleles (generating "read-only" contigs). BLAST analyses of "read only" contigs indicated that some known alleles (both *SRK* and paralogs) had been fragmented into multiple contigs. In such cases, contigs for a particular allele were combined, sequences sorted by barcode, and read numbers counted for each individual that contained a particular sequence type. Remaining "read only" contigs that did not show at least 80% similarity to *S*-related kinases from Genbank were not considered further. Final contigs were then sorted into putative "types": known *SRK* alleles, putatively new *SRK*-like variants, or known paralogs. Contigs assigned to *SRK* alleles whose dominance had been established previously (Prigoda et al., 2005; Goubet et al., 2012) were further sorted into the following classes: (1) A1, consisting of a single most recessive allelic specificity that has been found globally in *Arabidopsis* species (*SRK*01); (2) A2, dominant to all other classes; (3) A3, recessive only to class A2; and (4) B, recessive to all except A1 and showing high similarity to unlinked loci (*Aly*13-2 and *Aly*13-7). Contigs were also inspected for clustering of more than one named *SRK* allele from the database.

The next step was to subdivide variants within contigs into individual haplotypes, in order to test patterns of trans-specific polymorphism and to assess evidence for introgression between species. In our pilot study (Jørgensen et al., 2012) we recommended that only sequence variants present in at least 7% of the reads for an individual should be "counted" as true variants. However, in the full analysis, inspection of the contigs associated with particular alleles revealed very uneven read numbers both between individuals (ranging from a minimum of a single read to a maximum of 1,126 reads in the 465 individuals screened; average 344 ± 156) and across loci (i.e., *SRK* alleles and paralogs) within individuals. Low read numbers of particular alleles were also not directly proportional to the overall read numbers in the individual. The strict 7% threshold would have excluded some alleles that amplified in multiple individuals but were only present at low read numbers within individuals. A striking example was *SRK*01: it was fragmented across multiple contigs but when reassembled, it tended to be found at very low read numbers within individuals but was found across a wide range of individuals and showed population- and species-specific variants, as expected for a recessive allele (Billiard et al., 2007; Goubet et al., 2012). Many individuals showed <20 reads but the individuals that showed high read numbers (>100) tended not to show amplification of any other alleles, suggesting competition in the PCR when other alleles were present.

For haplotype calling, we thus also considered genotype calls at thresholds of at least 4% of reads and between 0 and 4% of reads. A problem with assessing such optimization strategies when including tetraploids is that there is not a robust basis for excluding individuals based on numbers of expected haplotypes. Although we could use diploids to determine thresholds of read numbers that minimized calling of more than two *SRK* alleles per individual and predicting homozygosity only for recessive alleles, this was confounded by the difficulties of predicting linkage of newly identified alleles (Charlesworth et al., 2003b; Prigoda et al., 2005). Tetraploids are expected to have up to four copies of *SRK* per individual but they can also contain multiple copies of recessive alleles (Mable et al., 2004), precluding extrapolating "confidence thresholds" based on diploids. We thus decided on a conservative threshold of at least 20 reads for a given haplotype to make relative comparisons among populations and species in the frequency of presence of particular variants. For reconstruction of evolutionary relationships among alleles, haplotypes present in <20 reads in a single individual and individuals with <200 total reads were excluded.

Statistical Analyses

To investigate whether there were differences in sequencing quality, detection biases, or real differences in frequency of sequence variants found we used generalized linear models to test whether the variation was significantly explained by ploidy, species or their interaction. Since multiple 454 runs were used for genotyping, we included barcoding tag number and lane as random effects, to account for any variation they explained. Analyses were conducted using JMP version 10.0 (SAS Institute, Incorporated).

Reconstructing Evolutionary Relationships Among Alleles

To establish phylogenetic relationships of newly identified alleles and to predict their dominance, we aligned the 454 sequences to the reference set (**Supplementary Data Sheet 1**) and reconstructed phylogenetic trees, using MEGA 7.0 (Kumar et al., 2016). We extracted consensus sequences for each haplotype of the *SRK*-like alleles identified and initially performed multiple alignments using the online version of Clustal Omega (Sievers et al., 2011) and then optimized by eye to establish positional homology and to set the correct reading frame to minimize stop codons, using Se-al version 2.0 (Rambaut, 1996) and McClade version 4.0 (Maddison and Maddison, 2000). To assess patterns of trans-specific polymorphism, if there was an exact match of a sequence to the reference database used for clustering, we named the haplotype "REF_HAP1" but if there was no exact match we retained the database allele (just named "REF"). We also added homologs from *A. lyrata*, *A. arenosa*, *A. halleri* and *A. thaliana* from Genbank for each specificity identified among the 454 samples (e.g., *AHASRK*04 and *ATH*-haplogroup A have been identified as homologs of *AlySRK*37; Bechsgaard et al., 2006). As implemented in MEGA, the best fitting substitution model was identified using ModelTest and then Maximum Likelihood was used to cluster sequences, using 1,000 bootstrap replicates. Due to the reticulate nature of evolution in this gene family, a strictly bifurcating evolutionary history is not expected but a tree-like representation is useful for identifying clusters of similar sequences. In previous studies, we have found that phylogenetic clustering is informative about dominance for Class A3 and B alleles but that Class A2 are paraphyletic based on alignments of approximately 900 bp of sequences in exon 1 of *SRK* (Prigoda et al., 2005). We thus used phylogenetic clustering to predict dominance of new specificities identified or known specificities for which dominance had not been established. We calculated genetic distances within and between dominance classes using both the best fitting substitution model and raw % similarity, using MEGA. We then mapped relative frequency of each haplotype in the four types of populations onto the tree, using Evolview in the Evolgenius package (He et al., 2016).

Testing the Accuracy of 454 Genotyping Using Segregation Analyses

We used the 454 pyrosequencing to genotype *SRK* from 11 families raised from crosses between tetraploid *A. lyrata* individuals whose grandparents had at least partially resolved *SRK* genotypes, in order to test segregation of alleles and as an additional test of reliability of the clustering thresholds set. Given the low read numbers found for *SRK*01, we established genotypes by a combination of allele-specific Sanger sequencing for this allele with the 454 sequencing for other alleles to compare segregation of alleles within families and to aid in excluding spurious allele calls. For a subset of these crosses, we performed controlled pollinations among all pairwise combinations of individuals, in order to test linkage of the variants identified to the SI phenotype and to predict dominance relationships (as in Mable et al., 2004).

Direct Sanger Sequencing of *SRK*01

To complement the 454 sequencing, we used targeted direct Sanger sequencing to resolve *SRK*01 genotypes to be able to investigate signatures of introgression of this recessive allele. We screened all individuals raised from the crosses between tetraploid *A. lyrata* individuals to aid in segregation analyses and a subset of individuals from the population survey to confirm haplotype calls and obtain more accurate frequencies of variants within and between individuals (**Table 1**).

We amplified products using an allele-specific primer (qtAl*SRK*01F: TCCTACATCATCGCAG) with the general reverse primer (SLGR: ATCTGACATAAAGATCTTGACC) that had been used for 454 sequencing. The 20 μL PCR reactions (using reagents from Invitrogen, Inc., Paisley, UK) consisted of 1 μL template, 2 μL 10x PCR buffer (Invitrogen Incorporated, Paisley, UK), 2 μL 10 mM dNTPs, 1 μL 50 mM MgCL2x, 0.2 μL 10 μM of each primer, and 0.2 μL *Taq* polymerase. The PCRs were run in MJ research thermocyclers using the following program: initial denaturing phase of 3 min at 94°C, 1 min annealing at 54°C, 2 min extension at 72°C; followed by 34 cycles of 30 s at 94°C, 30 s at 54°C, 2 min at 72°C; and a final extension step of 6 min at 72°C.

Individuals that showed amplification of products of the expected size (∼500 bp) were sent for sequencing to The GenePool in Edinburgh, using the reverse primer SLGR. Chromatograms were checked for base-calling errors using Sequencher 4.7 (Gene Codes Corporation, Ann Arbor, MI) and BLAST was used to confirm sequence identity.

Sequences were aligned using Sequencher, version 4.7 and heterozygous positions were recorded using IUPAC (International Union of Pure and Applied Chemistry) ambiguity codes. The phase of heterozygous positions was resolved by matching to variants found in the 454 sequencing and to homozygous sequences found in the Sanger sequencing. Genotypes predicted based on this process were then aligned to the specific 454 sequences for each individual. Species-specific variants were identified in diploids based on private haplotypes for the two species. We used the datamonkey server (www.datamonkey.org; Delport et al., 2010), which implements statistical tests associated with the programme HyPhy (Pond et al., 2005), to test for evidence of recombination using GARD (Genetic Algorithm for Recombination Detection; Pond et al., 2006). In addition, we manually inspected alignments for evidence of potential breakpoints and in such cases, aligned each "section" independently to the other haplotypes identified for a particular specificity. Where a putatively recombinant type showed similarity to two or more species-specific haplotypes in different regions of the sequence, they were classified as potentially introgressed. A minimum spanning network (Bandelt et al., 1999) was drawn using PopArt (Leigh and Bryant, 2015) to resolve the relationships among the *SRK*01 haplotypes.

Cloning and Sanger Sequencing of Longer *SRK* Alleles

As the 454 sequences were too short to be informative for future population genetics analyses and tests for selection, we used degenerate primers (**Supplementary Table 1**) to amplify longer products from tetraploid *A. lyrata* and *A. arenosa* sampled from the hybrid zone in the Wachau region of Austria (∼600 bp, also described in Ruiz-Duarte, 2012). We then used these products as seeds for the genome mining (see section Mining *SRK* Alleles From Genome Resequencing Data) to determine whether we could determine the genomic location of the "new" alleles found, as an indication of linkage to the *S*-locus.

Genomic DNA was extracted from three to four leaves from plants of tetraploid *A. lyrata* and *A. arenosa* individuals using a modified CTAB protocol (Doyle and Doyle, 1987). Degenerate primers known to amplify a number of different gene family copies and *SRK* alleles (Schierup et al., 2001) in *A. lyrata* and *A. halleri* (Forward: 13SeqF1, 5′-ccgacggtaaccttgtcatcctc-3′ and Reverse: SLGR, 5′-atctgacataaagatcttgacc-3′) were used (Charlesworth et al., 2000). Genomic DNA was mixed with a pair of primers, 10 μmol each, 4 μl of 5x buffer (ready-made), 50 mM MgCl2, 0.4 μl of 10 mM dNTP mixtures, 0.1 μl *Taq* DNA Polymerase (Mango *Taq*, Bioline). PCR amplification conditions were as follows: denaturation at 94°C for 2 min followed by 34 cycles of 94°C for 30 s, 50°C for 30 s, and 72°C for 30 s; a final extension at 72°C for 5 min.

PCR products were cloned into pGEM®-T Vector Systems (Promega Inc.). Colony PCR (20–30 colonies per individual) was conducted to test for inserts using SP6 and T7 primers, followed by Alu*I* digestion to identify clones carrying different putative *SRK* alleles. To avoid errors that might occur during PCR amplification and sequencing, a minimum of three independent clones with the same restriction profile were sequenced at the GATC BIOTECH facility. SeqMan software (DNASTAR, Inc) was used to clean and create consensus sequences.

We created separate alignments for each allele that was found both in the 454 and the Sanger sequencing by aligning the new sequences to references from Genbank and to the 454 sequences, in order to confirm shared specificity (**Supplementary Data Sheet 2**).

Mining *SRK* Alleles From Genome Resequencing Data

The 454 pyrosequencing data was not appropriate for determining presence and absence of paralogs because of: (1) the difficulty of distinguishing gene copies from new alleles at the *SRK* locus; and (2) amplification biases that made it difficult to set thresholds for reliability. Several known paralogs (*Aly*8, *Aly*9, *Aly*13-2/13-7) were expected to amplify with the primer set used. Polymorphic regions like the *S*-locus are known to be difficult to assemble in genome resequencing studies due to divergence from the reference genome (Mable et al., 2017) but we tested whether *de novo* assemblies from a genome resequencing study (Novikova et al., 2016) could be used to assess copy number of the *SRK*–related kinase gene family. We also attempted to pull out full-length sequences that spanned the *S*-domain (exon

1), transmembrane (exon 2) and kinase domains (exons 3-7) (Charlesworth et al., 2003a).

There are currently 28 fully resequenced genomes available from diploid and tetraploid *A. lyrata* and *A*. arenosa, from which we selected three or four individuals from each species and ploidy level to test whether we could obtain useful information on copy number and complete gene sequences. We used our paired end read data (Genbank SRR2040821, SRR2040822, SRR2040825, SRS945917, SRS1256176, SRS1256175, SRR2020827, SRR2040828, SRR2040829, SRR2040830, SRR2040791, SRR3111440, SRR3111441) and trimmed the reads for adapter contamination using cutadapt (Martin, 2011) and the respective adapter sequences. To obtain *SRK* alleles from these data we attempted two different approaches: mapping based and *de novo* assembly, on average we used ∼110 million paired end reads for the tetraploid accessions and ∼60 million reads for the diploid accessions corresponding to an average coverage of 20x.

In the initial mapping strategy we used as reference the *S*-locus region of the *SRK* locus on scaffold 7 of the MN74 reference genome (which was originally sampled from a North American outcrossing populations and has the S13 allele of the genes AL7G32720 = SCR, AL7G32730 = SRK, AL7G32710 = ARK3; Mable et al., 2017). Upon mapping we intended to extract reads that mapped to *SRK* and adjacent sequences in pairs and to perform a *de novo* assembly of these sequences only. In a first attempt we mapped reads using bwa (Li and Durbin, 2009). However, this approach did not yield any or an extremely low number of reads mapping to *SRK,* while adjacent regions were covered by the expected number of sequencing reads. Since bwa expects reads to have an identity of 90% or more to the reference and *SRK* alleles show much lower similarity (as little as 70% identity), we were not successful in mapping *SRK* reads to the reference. In a second attempt we used Next Gen Mapper (Sedlazeck et al., 2013), which only requires 65% of identity between read and reference. By this approach we were able to map reads to the *S*-locus including *SRK* but nevertheless a *de novo* assembly of these reads into complete or partial copies of the *SRK* locus failed.

We used CLC genomics workbench (https://www.qiagenbioinformatics.com/) to perform *de novo* assemblies using standard settings (automatic word and bubble size, minimum contig length 500 bp, reads were mapped back to contigs setting mismatch costs, insertion costs and deletion costs to 3 and length fraction as well as similarity fraction were set to 0.9) and the scaffolding option. Resulting scaffolds/contigs were indexed as BLAST libraries. We initially used FJ867321 (the *S*-domain from *AlySRK*30) to BLAST against these libraries to pull out sequences predicted to be *SRK* based on more than 50% coverage of the query sequence (filtered for low complexity, expect set to 10, word size to 11, match to 2, mismatch to −3, gap existence to 5, gap extension to 2). These hits were aligned to the first exon of AL7G32730 (*AlySRK*13 from the MN47 reference genome) to identify intron/exon boundaries and then trimmed if necessary. This approach yielded in total 66 sequences in the 13 accessions analyzed (**Supplementary Table 10**). Therefore, our BLAST search also must have identified other S-domain encoding genes besides *SRK*.

In order to obtain an overview on the presence of *S*-domain encoding genes we performed another BLAST search using the first exon of the MN47 *SRK* against the MN47 reference genome. This search revealed five genes encoding proteins that have an *S*-domain (*AL7G32730* = *SRK*, *AL7G32710* = *Aly8*, *AL6G48380* = *Aly3*, *AL3G23610* = *Aly*9, *AL2G23090* = *Aly10.2*). From this result we expected that our contigs identified in the 13 resequenced accessions should have their best BLAST hit with one of these five loci. So, we aligned the 66 contig sequences to the first exon of the MN47 *SRK* and trimmed them in length to the first exon. Then we performed a blast search of the 66 trimmed sequences against the MN47 reference genome. All of the 66 sequences had their best blast hit with one of the five loci we had identified beforehand. Typically hits for *AL7G32710*, *AL6G48380*, *AL3G23610*, *AL2G23090* showed a very small E-value and a high score while *AL7G32730* hits were characterized by a lower score and E-value due to the lower conservation for alleles of this locus.

We initially used the BLAST results to predict similarity to known *SRK* alleles and related receptor kinase gene family members available in Genbank for each of the contigs. However, since we had identified potentially new variants in this study, we also aligned sequences pulled out from the resequenced genomes to our reference database and to the sequences found using 454 and the longer Sanger sequences to confirm sequence identity (**Supplementary Data Sheet 2**; **Supplementary Table 12**). We used clustering in phylogenetic trees (reconstructed using Maximum Likelihood in MEGA 7.0) to predict *SRK* specificity and to determine presence/absence of other members of the gene family.

One of the paralogs (*Aly*9) is known to amplify in all *A. lyrata* individuals that have been tested using PCR-based screening (Mable, personal observation). We thus used identification of this locus as a control for whether it was likely that the genome-mining approach could be reliably used to detect copy number variation in highly polymorphic gene families. The approach described initially only identified this locus in three of 12 genomes so we trialed another approach, using the sequences in **Supplementary Data Sheet 2**, along with the 66 contigs originally identified to BLAST the *de novo* assemblies for each genome. This resulted in an additional 102 contigs, which then were aligned back to the reference database and identities confirmed using cluster analysis. In this analysis *Aly*9 was resolved for all individuals and more complete genotypes were obtained for *SRK* and the other paralogs screened, so only the results from this final analysis are presented.

RESULTS AND DISCUSSION

Objectives 1 and 2: Diversity and Allele Sharing of *SRK* in Diploids and Tetraploids

After filtering and assigning variants to alleles based on sequence similarity and predicting dominance classes and linkage to the SI phenotype based on phylogenetic clustering, we identified 107

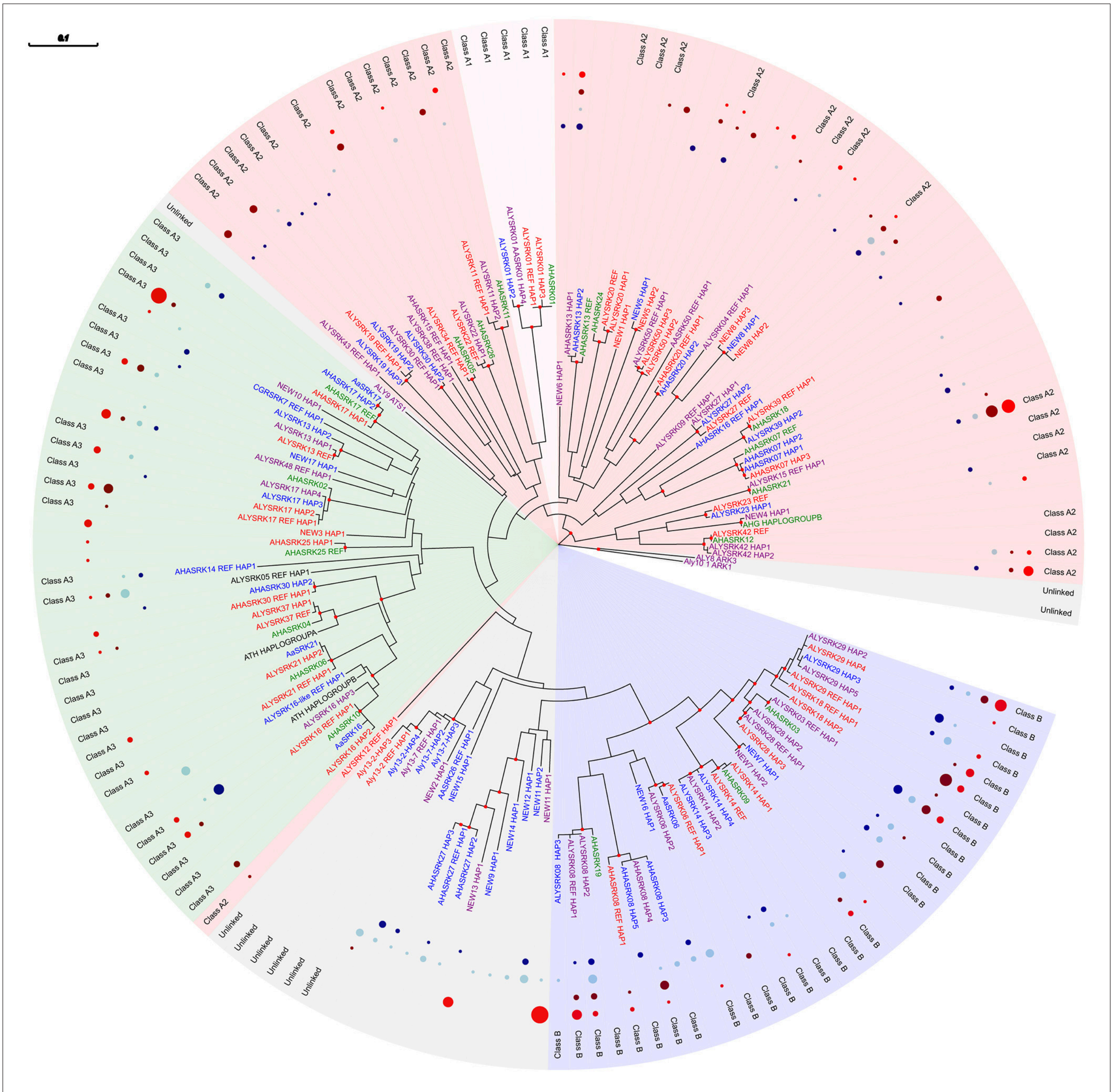


FIGURE 1 | Maximum likelihood tree based on SRK-like sequences resolved through 454 pyrosequencing, reconstructed using MEGA 7 under an HKY85 model of evolution with rate heterogeneity modeled under a gamma distribution and with proportion of invariant sites estimated. Bootstrap proportions above 70% are indicated as filled circles on nodes. The tree was rooted with the unlinked paralogs *Aly*8 (*Ark*3 in *A. thaliana*) and *Aly*10.1 (*Ark*1 in *A. thaliana*). Alleles for each *SRK* specificity are assigned to a dominance class based on previous studies of *A. lyrata* (Prigoda et al., 2005) and *A. halleri* (A1 = yellow; A2 = red; A3 = green; B = blue; unlinked = gray); new alleles or previously identified alleles where dominance has not been confirmed are colored according to the class predicted by their position in the tree. Tip labels are colored according to the species in which they were found in the 454 sequences (lyrata = red; lyrata+arenosa = purple; arenosa = blue) or the origin of the reference allele in cases where there was no exact match (halleri = green; thaliana = black). Also shown is the frequency of a particular haplotype in each of the four groups compared (diploid arenosa, A2x = dark blue; tetraploid arenosa, A4x = light blue; diploid lyrata, L2x = dark red; tetraploid lyrata, L4x = light red). Due to the high number of haplotypes but low read numbers for *AlySRK*01 and the unlinked loci *Aly*13-2 and *Aly*13-7, only a subset of haplotypes are included and frequencies are not indicated.

haplotypes (unique sequence variants) that could be grouped into 63 potential alleles (specificities) that were at least 80% similar to *SRK* (**Figure 1**; **Supplementary Table 2**). Seventeen were potentially new specificities that were <90% similar to the *A. lyrata*, *A. halleri* or *A. arenosa* reference sequences included (**Supplementary Table 3**). However, seven of these new variants were predicted not to be linked to SI based on phylogenetic clustering and so could represent other members of the gene family. All of the new potentially unlinked alleles were found in diploid and/or tetraploid *A. arenosa*, with two of them also occurring at high frequency in L4x populations but only a single L2x individual sharing one of the new unlinked alleles with A4x individuals. The new alleles predicted to be linked to SI were distributed more evenly among the two species.

When accounting for variation due to lane and tag as random effects using generalized linear models, we found no evidence for significant differences between species or ploidy levels or their interactions in terms of number of reads, total number of contigs resolved (indicative of the wider gene family), the number of *SRK*-like alleles (i.e., variants showing at least 80% similarity to known *SRK* sequences, so including unlinked alleles), or the number of alleles or haplotypes per individual predicted to be linked to *SRK* (**Supplementary Table 4**). There was a significant interaction between ploidy and species in the proportion of contigs resolved that were at least 80% similar to *SRK* (i.e., more reads were *SRK*-like than similar to other members of the gene family), with a significantly higher proportion in tetraploids compared to *A. arenosa* diploids but no significant difference compared to *A. lyrata* diploids. Since the primers used were developed based on variation within *A. lyrata* (Schierup et al., 2001; Charlesworth et al., 2003a, 2006), this could be an indication that not all *SRK*-like alleles were amplified for *A. arenosa* due to variation in the primer regions, resulting in resolution of more spurious contigs due to non-specific amplification. However, overall, there was very little evidence that tetraploids were fundamentally different to diploids in terms of sequence quality or the ability to resolve variants.

The 200 bp sequences produced similar resolution in phylogenetic clustering as previous studies using 600 bp (Tedder et al., 2011) and resulted in consistent patterns of polymorphism expected for dominant and recessive alleles at *SRK*. Examination of relative frequency distributions also generally met theoretical expectations but indicated no obvious differences in diversity between ploidy levels. There was extensive variability in relative frequencies of each haplotype, with some being restricted to certain species or populations and some being found across both species and ploidy levels (**Figure 1**; **Supplementary Table 2**). We predicted that there should be highest interspecific sharing of individual haplotypes among tetraploids due to their known introgression (Schmickl, 2009; Schmickl et al., 2010; Jørgensen et al., 2011; Schmickl and Koch, 2011) but also because they can maintain more allelic copies within individuals. We found that 23 haplotypes were shared between A4x and L4x compared to 12 between A2x and L2x, including seven that were shared among all four population types (**Supplementary Table 2**). Sharing between the two types of tetraploids was similar to that among ploidy levels within species (24 among *A. lyrata* and 22 among *A. arenosa*). The highest number of private haplotypes was also found for diploids: 19 for A2x and 15 for L2x, compared to 12 for A4x and 8 for L4x. These results are consistent with predicted patterns of introgression among the tetraploids in northeastern Austria (Wachau region and Forealps; Schmickl et al., 2010; Jørgensen et al., 2011; Schmickl and Koch, 2011).

Although it is difficult to separate increased transpecific polymorphism from this introgression, we found some evidence that there might be more differences in selection pressure or demographic history between species than between ploidy levels. Plotting allele frequency distributions for each ploidy and species combination demonstrated an excess of intermediate frequency alleles in both diploid and tetraploid *A. arenosa* (**Figure 2**), as expected for a locus under balancing selection (Mable and Adam, 2007). However, the pattern was more skewed toward low frequency alleles in *A. lyrata*, particularly in tetraploids. In North American populations of *A. lyrata*, a difference in allele frequency spectrum for *SRK* was found between inbreeding and outcrossing populations (Mable and Adam, 2007) but the latter showed more similar patterns as those observed for *A. arenosa* in this study. Since shifts toward intermediate frequencies are also expected for population bottlenecks (Luikart et al., 1998), it is possible that in particular diploid *A. arenosa* experienced a larger decline in population numbers since the past glaciation. What was striking in the current study was that tetraploids did not have a dramatically higher number of alleles or haplotypes within populations or alleles or haplotypes per individual than diploids, regardless of dominance class (**Table 2**). Furthermore, for neutral genes, there is a steep gradient of increasing genomic contribution of A4x found within introgressed *A. lyrata* along a transect in the hybrid zone (Schmickl, 2009; Schmickl et al., 2010; Jørgensen et al., 2011; Schmickl and Koch, 2011; Hohmann et al., 2014; Muir et al., 2015) but this is not reflected in the *SRK* distribution; i.e., *SRK* are more mixed than would be predicted based on neutral patterns, as might be expected under balancing selection. This suggests that tetraploids are not fundamentally different from diploids in their capacity for maintaining diversity of *SRK*, as suggested previously from segregation analyses within tetraploid families based on crosses involving one of the same tetraploid populations studied here (L4_AUT2) and a tetraploid population from Aggsbach, Austria (Mable et al., 2004).

Consistent with theory (Billiard et al., 2007), recessive alleles in diploids have been demonstrated to occur at higher frequency, to show shallower branch lengths in phylogenetic analyses, and more extensive polymorphism within specificities than dominant alleles (Llaurens et al., 2008, 2009; Castric et al., 2010; Vekemans et al., 2011; Goubet et al., 2012). In our study, Class B alleles (recessive to A2 and A3 classes) showed lower intraclass polymorphism (13% average pairwise sequence divergence, compared to 25% for Class A2 and 15% for Class A3) but more haplotypes per allele than the two dominant classes (2.56 ± 1.33 compared to 1.59 ± 0.75 in Class A2 and 1.56 ± 0.89 in Class A3, **Table 2**) and there was high divergence between classes (26–29%; **Table 3**). The paralogous locus identified in previous studies that is similar to class B alleles (*Aly*13-2) showed similar within locus variation (13%) as for class B alleles and lower divergence from class B than the other dominance classes (16% compared to at least 27% to the others). There was a higher proportion of alleles restricted to only one of the species among the dominant (29% for Class A2 and 50% for Class A3) than recessive (20% for Class B) alleles but a majority of the unlinked alleles (67%) were only found in *A. arenosa* (**Supplementary Table 2**). Thirteen alleles were found only in

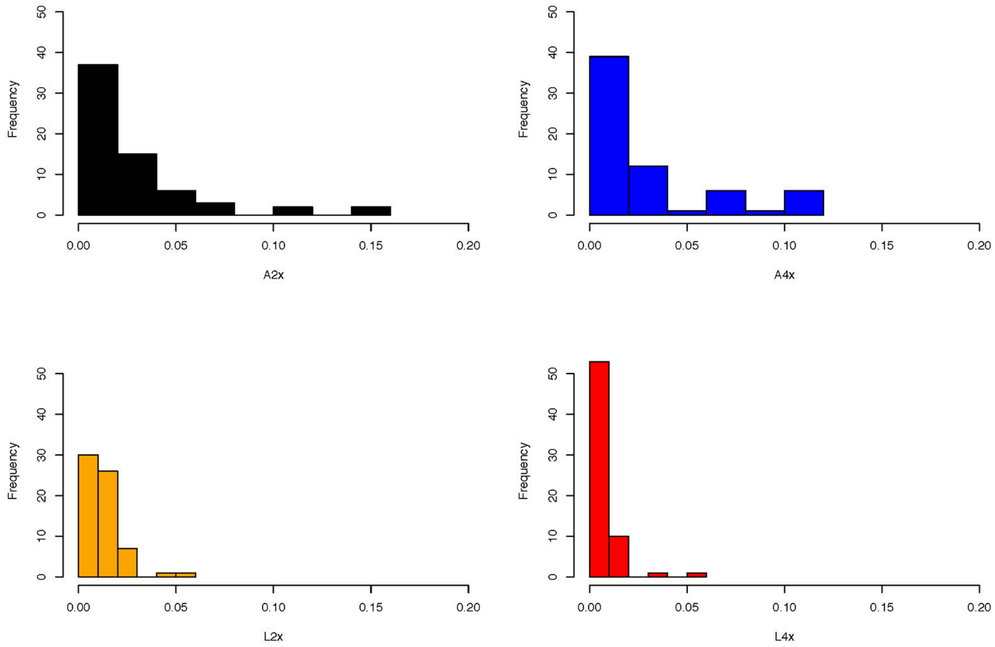


FIGURE 2 | Allele frequency distributions of *SRK* alleles identified in diploid and tetraploid populations of *A. lyrata* and *A. arenosa* using 454 pyrosequencing. Note that there appears to be an excess of intermediate frequencies in *A. arenosa* (A2x = diploids; A4x = tetraploids), with more of a skew toward low frequency alleles in *A. lyrata*, particularly in tetraploids.

tetraploids, but none were Class B and only four (three Class A2 and one Class A3) were shared between the two species. Thus, results were consistent with the increased trans-specific polymorphism expected for recessive alleles at a locus under balancing selection (Billiard et al., 2007; Llaurens et al., 2008; Castric et al., 2010; Goubet et al., 2012).

Overall, these results suggest that tetraploids do not show increased mate availability due to an increase in *S*-locus repertoire but instead might be constrained by the potential mate limitation caused by having "too many" S-alleles. This is similar in theory to expectations for immune genes in animals, where an optimal number of alleles has been suggested as conferring higher fitness than maximizing allelic diversity (Reusch et al., 2001; Aeschlimann et al., 2003; Wegner et al., 2003; Kalbe et al., 2009). The high allele sharing among ploidy levels precluded testing of whether there is relaxed balancing selection acting in tetraploids but this was not suggested by the site frequency distributions, which suggested a stronger species than ploidy effect. Nevertheless, there are some important caveats to consider in the interpretation of these results, due to particular challenges when working with this type of gene families (see Challenges below).

In the crosses between tetraploid *A. lyrata* individuals, we found the same three *SRK*01 haplotypes using both 454 and targeted Sanger sequencing (haplotypes 1, 2, and 3). This allowed us to test the accuracy of the 454 genotyping despite the low read numbers for *SRK*01 and provide more complete data for segregation analyses. For 50% of the individuals identical genotypes were predicted using the two approaches, with 14% testing negative for the allele-specific PCR but positive using 454, compared to 10% showing the opposite pattern (**Table 4**). Different haplotypes were predicted by the two methods only for a single individual. However, the direct sequencing was more sensitive, resolving heterozygotes in 24% of the individuals that were predicted to be homozygous based on the 454 sequencing (compared to only 2% showing the opposite pattern). Segregation of *SRK*01 genotypes in the crosses confirmed previous predictions (Mable et al., 2004) that tetraploids could harbor multiple copies of haplotypes for this recessive specificity (**Table 5**). These data were then combined with segregation of the haplotypes resolved using 454 pyrosequencing (**Supplementary Table 5**). After excluding 454 alleles not present in the parents, the majority of individuals showed four or fewer expected haplotypes. Comparison of segregation of predicted genotypes with self-incompatibility phenotypes (**Figure 3**; **Supplementary Tables 6**, **7**), confirmed linkage of two alleles previously tested in other crosses (*SRK*16 and *SRK*29) and one that had been identified in the grandparents but had not been deposited to Genbank (*SRK*48). However, the segregation analyses suggested that not all alleles were detected by 454 and suggested that the stringent filtering in some cases omitted alleles that must have been present based on the incompatibility phenotypes.

Challenge: Filtering Decisions for Clustering

Despite recommendations from our pilot study that a threshold of 90% similarity would be appropriate for clustering (Jørgensen et al., 2012), our analyses of the full dataset suggested that

TABLE 2 | Distribution of alleles (A) and haplotypes (B) across diploid (A2x, L2x) and tetraploid (A4x, L4x) populations (POP) for different predicted dominance classes (A2 and A3 are dominant to B), excluding Class A1, which is represented only by *SRK*01; read numbers were too low to be certain about presence or absence for that allele.

(A)

		N ALLELES				**N ALLELES/IND**			
POP	**N IND**	**A2**	**A3**	**B**	**ALL**	**A2**	**A3**	**B**	**ALL**
A2x	75	16	9	10	35	0.21	0.12	0.13	0.47
A4x	77	15	11	13	39	0.19	0.14	0.17	0.51
L2x	70	18	7	8	33	0.26	0.10	0.11	0.47
L4x	102	14	11	10	35	0.14	0.11	0.10	0.34

(B)

		N HAPLOTYPES				**N HAPLOTYPES/IND**			
POP	**N IND**	**A2**	**A3**	**B**	**ALL**	**A2**	**A3**	**B**	**ALL**
A2x	75	19	9	18	46	0.25	0.12	0.24	0.61
A4x	77	15	11	19	45	0.19	0.14	0.25	0.58
L2x	70	20	9	13	42	0.29	0.13	0.19	0.60
L4x	102	16	14	17	47	0.16	0.14	0.17	0.46
Total		70	43	67	180				

Alleles that did not appear to fall under any of the known dominance classes are not included, as they were predicted to be unlinked to the SI phenotype. Also shown is the number of alleles or haplotypes per individual.

TABLE 3 | Percent sequence divergence within and between dominance classes, for alleles identified using 454 sequencing.

Class	**B**	**A1**	**A2**	**A3**	***Aly*13-2**
B	0.129				
A1	0.282	0.032			
A2	0.287	0.259	0.253		
A3	0.279	0.265	0.277	0.151	
*Aly*13-2	0.159	0.284	0.272	0.267	0.129

Divergence within classes is shown on the diagonal. An unlinked locus that shows polymorphism among haplotypes (Aly13-2) is included for comparison.

TABLE 4 | Proportion of individuals that tested positive for *SRK*01 specificity using direct Sanger and 454 sequencing, indicating the population (A2x = diploid *A. arenosa*; A4x = tetraploid *A. arenosa*; L2x = diploid *A. lyrata*; L4x = tetraploid *A. lyrata*), sample sizes (N-direct, N-454) and % of individuals that tested positive for *SRK*01 in each.

Population	**N-direct**	**% *SRK*01-direct**	**N-454**	**% *SRK*01-454**
A2x	44	20.5	78	66.7
A4x	65	44.6	87	78.2
L2x	57	31.6	76	55.3
L4x	79	41.8	115	61.7
Total	245	36.6	191	59.2

a single threshold may not be appropriate for gene families that include different levels of divergence among classes or copies; for example, in relation to dominance (Prigoda et al., 2005). In our study, BLAST analysis of "read only" contigs demonstrated that some known alleles were fragmented across multiple contigs. For recessive alleles (Class B, *SRK*01) and unlinked loci (*Aly*9, 13-2 and 13-7), combining contigs resulted in mixtures of haplotypes from different alleles (specificities), making it challenging to assign sequence variants to alleles. While several dominant alleles (*Aly*16, *Aly*30, and *Aly*42) also showed fragmentation, there was no ambiguity in assigning sequence variants to alleles. Resolving recessive alleles into unique contigs thus required more manual manipulation and sorting of variants into haplotypes. Since recessive alleles also had on average more haplotypes per allele (2.44 ± 1.42) than dominant alleles (1.57 ± 0.74 for A2; 1.56 ± 0.89 for A3) (**Supplementary Table 2**), read numbers per haplotype were often lower, which made setting a single threshold for reducing spurious genotyping difficult.

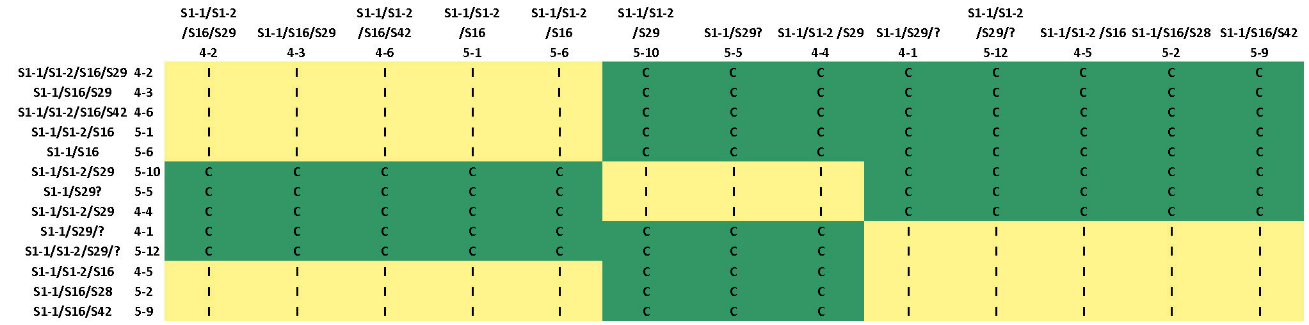
Challenge: Amplicon Based Errors and Biases

From previous studies we anticipated that the single most recessive allele, *SRK*01, would be present at high frequency and would show a higher number of haplotypes than other specificities (Billiard et al., 2007; Castric and Vekemans, 2007; Llaurens et al., 2008; Castric et al., 2010; Goubet et al., 2012; Vekemans et al., 2014). In our 454 data, *SRK*01 was present in all populations surveyed and we identified 15 unique variants that were present in more than one individual; however, read numbers tended to be very low (often with <10 reads per individual) and fell well below the thresholds set for considering "real" presence of a given haplotype used for other loci for most individuals. Although multiple haplotypes differing by a single or few bp are expected for recessive alleles (Castric and Vekemans, 2007), the low read numbers made it difficult to distinguish PCR errors from actual polymorphism. High read numbers were found for some individuals, but they tended to show the presence of few other sequence variants. In addition, several known paralogs that should be present in all individuals (*Aly*8, *Aly*9; Charlesworth et al., 2003b) were expected to amplify with the primer set used but this was very inconsistent. *Aly*9 was present in the majority of individuals but read numbers varied dramatically from 0.5 to 92% of the total reads in an individual. There was a significant difference in the proportion of reads that were *Aly*9, with *A. arenosa* tetraploids showing a higher proportion than both diploids, which showed a significantly higher proportion than *A. lyrata* tetraploids (**Supplementary Table 4**). Whether this is due to an amplification bias or expansion of the gene family is difficult to distinguish. For *Aly*8, only 41/460 sequenced individuals showed any amplification and most were present at only low read numbers (maximum 15%). We thus could not assess presence or absence of other members of the gene family based on the 454 sequencing or use the paralogs to make inferences about introgression in the tetraploids to avoid the confounding effects of balancing selection. Even after correcting for chimeras, there was some evidence for recombination in some of the specificities showing polymorphism among populations (e.g., *SRK*01, some of the class B alleles) but this was difficult to

TABLE 5 | Segregation of *SRK*01 genotypes within families raised from crosses between tetraploid *A. lyrata* individuals, as determined by direct Sanger sequencing; the number of individuals where a particular genotype was found is indicated in parentheses.

Cross	N	Genotypes					
A1XB4	8	1-1/1-2 (4)	1-1 (4)				
A1XC1	7	1-1 (4)	1-1/1-2 (3)				
A1XC2	15	1-1 (7)	1-1/1-2 (8)				
A1XC4	7	1-1 (5)	1-1/1-2 (1)	no SRK01 (1)			
A1XE3	10	1-1/1-2 (8)	1-1 (2)				
C3XE7	7	1-1/1-3 (5)	1-1 (2)				
C3XE8	9	1-1/1-2 (6)	1-1 (1)	no SRK01 (2)			
E6XC1	10	1-1/1-3 (6)	1-1 (4)				
E6XE1	7	1-1 (3)	1-3 (1)	1-1/1-3 (2)	no SRK01 (1)		
E8XC3	9	1-1 (1)	1-1/1-3 (2)	1-1/1-2/1-3 (1)*	no SRK01 (5)		
E8XE11	6	1-1 (1)	1-3 (1)	1-1/ 1-3 (1)	1-1/1-2 (1)	1-1/1-2/1-3 (1)	no SRK01 (1)
E8XE6	8	1-1/1-3 (3)	1-1 (1)	no SRK01 (4)			

Complete segregation of haplotypes found using 454 sequencing, combined with this genotyping is detailed in ***Supplementary Table 5****. *homozygous for SRK01 in 454 sequencing.*



	S1-1/S1-2 /S16/S29 4-2	S1-1/S16/S29 4-3	S1-1/S1-2 /S16/S42 4-6	S1-1/S1-2 /S16 5-1	S1-1/S1-2 /S16 5-6	S1-1/S1-2 /S29 5-10	S1-1/S29? 5-5	S1-1/S1-2 /S29 4-4	S1-1/S29/? 4-1	S1-1/S1-2 /S29/? 5-12	S1-1/S1-2 /S16 4-5	S1-1/S16/S28 5-2	S1-1/S16/S42 5-9
S1-1/S1-2/S16/S29 4-2	I	I	I	I	I	C	C	C	C	C	C	C	C
S1-1/S16/S29 4-3	I	I	I	I	I	C	C	C	C	C	C	C	C
S1-1/S1-2/S16/S42 4-6	I	I	I	I	I	C	C	C	C	C	C	C	C
S1-1/S1-2/S16 5-1	I	I	I	I	I	C	C	C	C	C	C	C	C
S1-1/S16 5-6	I	I	I	I	I	C	C	C	C	C	C	C	C
S1-1/S1-2/S29 5-10	C	C	C	C	C	I	I	I	C	C	C	C	C
S1-1/S29? 5-5	C	C	C	C	C	I	I	I	C	C	C	C	C
S1-1/S1-2/S29 4-4	C	C	C	C	C	I	I	I	C	C	C	C	C
S1-1/S29/? 4-1	C	C	C	C	C	C	C	C	I	I	I	I	I
S1-1/S1-2/S29/? 5-12	C	C	C	C	C	C	C	C	I	I	I	I	I
S1-1/S1-2/S16 4-5	I	I	I	I	I	C	C	C	I	I	I	I	I
S1-1/S16/S28 5-2	I	I	I	I	I	C	C	C	I	I	I	I	I
S1-1/S16/S42 5-9	I	I	I	I	I	C	C	C	I	I	I	I	I

FIGURE 3 | Segregation analysis for the cross A1 × C2, based on combined genotypes from the direct sequencing of *SRK*01 and from 454 pyrosequencing based on genotyping of other alleles found in the crosses. The cross was between two different tetraploid *A. lyrata* populations in Austria that were thought to be outside of the hybrid zone with *A. arenosa*. The predicted genotype of the donor is indicated along the top row and that of the recipient in the column to the left. Incompatible cross combinations are indicated with an I (and shaded yellow), compatible combinations with C (and shaded green). Comparison of segregation of *SRK* haplotypes with the phenotype suggests that: S16 is expressed with all other haplotypes; it appears to be codominant with S42 in stigmas but recessive in pollen. S29 is recessive to S16. Individuals 4-1 and 5-12 must have an allele that has not been identified because they show different patterns of compatibility than 5-10, 5-5, and 4-4, which also only have S1-1 and S29. S28 was only found in three individuals and only one (5-2) was included in the analyses shown here; based on the 454 genotyping it can be difficult to distinguish S18, S28, and S29 so this could be an error in assignment. Individual 5-5 must have S29 but it was not detected in the 454 analyses.

distinguish from PCR recombinants, particularly with only 200 bp of sequence.

Challenge: Assessing the Accuracy of Genotyping

Although arguably more problematic for 454 pyrosequencing than for more recently developed approaches due to tag switching of barcodes, which we previously found could occur for up to 7% of samples (Jørgensen et al., 2012) and has been reported in other studies (Carlsen et al., 2012), the biggest challenge was deciding on thresholds and criteria for assessing accuracy of genotyping and efficiency of filtering strategies. The 200 bp sequences resolved were useful for assessing haplotype diversity within alleles, identifying putatively new alleles, predicting dominance based on phylogenetic clustering, and the distribution of allele and haplotype frequencies among populations. The results also generally fit with theoretical predictions. However, there was less certainty for determining individual genotypes; the crosses, for example, included more alleles than should have been present in some individuals, including alleles that were not identified in the parents (**Supplementary Table 5**). The haplotype frequencies indicated in **Figure 1** and **Supplementary Table 2** are thus based on a conservative threshold of at least 20 reads per individual but this likely underestimates patterns of haplotype sharing across populations and species. Nevertheless, an advantage of studying gene family evolution in SI genes over comparable systems like the MHC in vertebrates is that linkage of each new variant could be tested by segregation analyses to a known phenotype (Schierup et al., 2001; Mable et al., 2003, 2004; Prigoda et al., 2005). In our study, the low amplification of *SRK*01, which we otherwise knew from Sanger sequencing based genotyping of the parents should have multiple variants within families, precluded confidence in segregation analyses based only on the 454 data. However, targeted Sanger sequencing for this allele aided in interpretation of the segregation analyses. Unfortunately, as we performed crosses before the 454 sequencing, we could not test linkage of all new variants found to the SI phenotype. It was also not feasible to determine when unlinked alleles were amplified based on the presence of "too many" haplotypes.

Objective 3: Introgression of *SRK* Alleles

For the population survey, the 454 genotyping identified 22 *SRK*01 variants. Using targeted amplifications and Sanger sequencing we identified 24 haplotypes. All of these but seven had been found using the 454 pyrosequencing, but including five that matched the 454 sequences but had additional polymorphisms outside of the shared sequence region (indicated by distinct letters after the haplotype name; **Supplementary Tables 8**, **9**). However, only 11 of the 22 variants found by 454 sequencing were confirmed by direct sequencing and there was a higher proportion of PCR positive results among the 454 than the Sanger sequences (**Table 4**).

Using the diploids as a guide, we identified "arenosa" and "lyrata" specific haplotypes, as well as three that appeared to be recombinants between species-specific variants (haps 7, 8, and 10; **Supplementary Data Sheet 3**), two of which were identified from a single A4x population that was predicted to be introgressed (A4X_AUT1, from Kerhnoff; Schmickl, 2009). Although analyses using GARD in the HYPHY package did not find statistical evidence for recombination breakpoints, this might have been because of the short tracts of introgression. The minimum spanning network indicated that haps 7 and 8 did in fact fall between species-specific clusters whereas hap10 was on a tip in the *A. arenosa* part of the network (**Figure 4**). What is striking is that reticulation in the network involved primarily *A. arenosa* tetraploids and that diploids had a lower diversity of *SRK*01 haplotypes compared to tetraploids. There was also some haplotype sharing among tetraploids but not between the diploids. Since the crosses established that individual tetraploids could harbor up to three different *SRK*01 haplotypes and many were heterozygous for two, this higher diversity among tetraploids could be because *SRK*01 is effectively neutral and so could accumulate more mutations in tetraploids because of the higher copy number maintained (Mable et al., 2004). Crossing data suggest that *SRK* is functional in individuals sampled from the hybrid zone (Ruiz-Duarte, 2012), but it is also possible that selection pressure to maintain restricted recombination in the *S*-locus region (Charlesworth et al., 2006) would be relaxed with the increased copy number in tetraploids. Moreover, introgression of recessive alleles between *A. lyrata* and *A. halleri* has been found in diploids (Castric et al., 2010), suggesting that hybridization might disrupt linkage. Although the crosses we performed only included tetraploid *A. lyrata* from outside of the known hybrid zone, two individuals in one family were self-compatible (**Supplementary Table 6**). It is thus also possible that increased recombination at the *S*-locus occurs with spontaneous loss of SI in some individuals.

The presence of *A. arenosa* like haplotypes in two of the *A. lyrata* tetraploid populations and the most frequent *A. lyrata* haplotype (hap1) in most *A. arenosa* populations from the hybrid zone (**Table 4**, **Supplementary Table 9**) could suggest more recent and secondary hybridization while the introgressed haplotypes (i.e., those that appeared to be recombinants between the species-specific variants) could reflect older events. One *A. arenosa*-like haplotype (hap2) was found in an *A. lyrata* tetraploid population in the Northeastern Austrian Forealps (L4_AUT4 from Lilienfeld), and in the crosses, which involved individuals from two peripheral *A. lyrata* populations (L4_AUT2 from Mödling and L4_AUT5 from Rauheneck Ruin, near Baden). This could suggest undetected hybridization within these "pure" populations, as also suggested by whole-genome data (Hohmann and Koch, 2017). While these results fit with expectations based on predicted patterns of hybridization in tetraploid populations from Austria (Schmickl, 2009; Schmickl et al., 2010; Schmickl and Koch, 2011; Muir et al., 2015), there are similar caveats about the use of PCR-based genotyping as raised for the 454 sequences, as described below.

Challenge: PCR Based Approaches to Genotyping

Overall, there was not much consensus between the *SRK*01 genotypes resolved using 454 and direct sequencing. While the crosses demonstrated that the latter was more sensitive to detect heterozygotes when products were amplified, the population survey revealed a potential bias against amplifying variants found in A2x populations. A much lower proportion of individuals from these populations tested positive than from other populations, and many of the haplotypes found using 454, but not direct sequencing, were from A2x populations. This potential bias reduced the sample sizes that could be used to classify haplotypes showing species-specific presence. In the segregation analyses (**Table 5**), two individuals had all three *SRK*01 haplotypes segregating in the parents: one individual didn't show presence of other alleles expected in the parents based on the 454 sequencing but showed some unexpected alleles; the other individual showed more than four expected haplotypes (**Supplementary Table 5**). Thus, we cannot rule out contamination. Moreover, interpretation of introgressed haplotypes could have been confounded by PCR-based recombination but they were found only in a stabilized hybrid population. Moreover, some haplotypes were only resolved from direct sequences of heterozygotes; in those cases cloning would be required to absolutely confirm the full range of haplotypes present. We had originally intended to also test the utility of other polymorphic members of the gene family (e.g., *Aly*9); however, since there were even more haplotypes predicted by the 454 sequencing but separated by fewer variants (data not shown), there would have been too much reliance on accurately identifying singletons.

Objective 4: Copy Number Variation in the *SRK*-Related Gene Family

Clustering of contigs resolved from the *de novo* assembly approach to genome mining of our database of *SRK* and its paralogs (i.e., all unique variants found using the 454 pyrosequencing, targeted sequencing of *SRK*01, cloning of longer

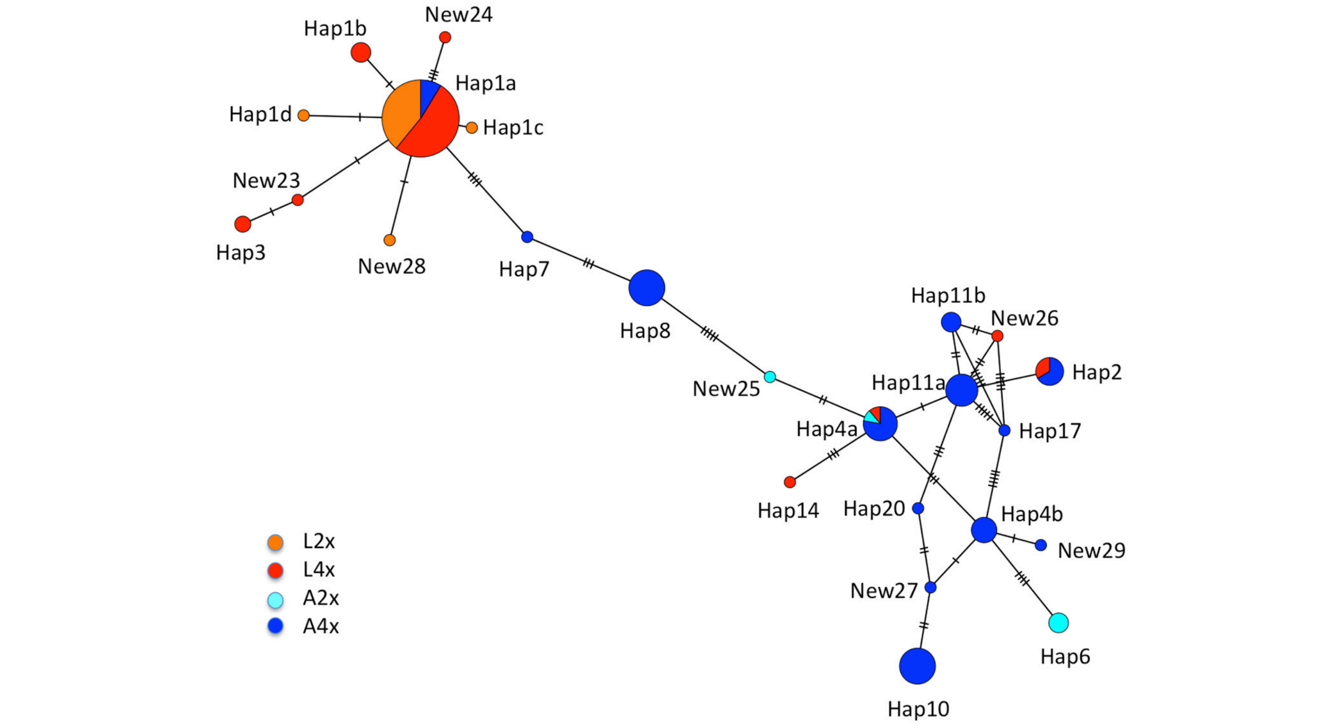


FIGURE 4 | Minimum spanning network for *SRK*01 haplotypes resolved using direct Sanger sequencing. Circles are drawn proportionately to the frequency of the haplotype and colored by relative frequency in each population type: A2x = light blue; A4x = dark blue; L2x = orange; L4x = red. Vertical bars on the connecting branches indicate the number of nucleotide substitutions separating haplotypes. Haplotypes 7, 8, and 10 were predicted to be recombinants between "arenosa" and "lyrata" specific sequences; haps 7 and 8 appear intermediate between the two clusters whereas hap10 is on a tip in the "arenosa" part of the network. NEW25 also appears intermediate and so could be another introgressed haplotype but it was only found in a single individual. Note that extensive reticulation was found predominantly among sequences found in A4x populations and that there is less variation among haplotypes restricted to diploids than those found in tetraploids.

products using degenerate primers, and additional sequences available in Genbank) was used to uncover receptor-like kinases from published diploid and tetraploid genome sequences (**Supplementary Table 10**). This resulted in identification of 1-2 predicted *SRK* alleles in the diploid and 1-4 in the tetraploid accessions for both species among the 13 short read sets screened (**Table 6**; **Supplementary Table 11**). In total 29/177 contigs were assigned as *SRK*, but 12 of these would have been mis-assigned based only on BLAST (**Supplementary Table 10**). *Aly*13-2-like sequences were pulled out in seven accessions, but would have been classified as *SRK* based only on BLAST (**Table 6**; **Supplementary Table 10**). This locus is not present in all individuals, so copy number variation is expected (Mable et al., 2017). Other alleles predicted to be unlinked to the SI phenotype were also resolved by the clustering analysis but none of these would have been assigned as *SRK*-like based on BLAST (**Table 6**; **Supplementary Table 10**). One published allele (*AlySRK*32) whose phylogenetic position and dominance have not been resolved in previous studies was pulled out from six accessions; based on the length of its branch to other *SRK* sequences, it has been predicted to be unlinked to the SI phenotype (Tedder et al., 2011). *AlySRK*47 (found in four of the accessions) is also predicted to be unlinked, based on its phylogenetic position relative to linked sequences. The other four paralogs tested were present in all accessions, except for one diploid *A. lyrata* that lacked *ARK*3 (*Aly*8). Since this latter locus is tightly linked to *SRK* in some specificities and shows high polymorphism (Kusaba et al., 2001; Charlesworth et al., 2003b; Guo et al., 2011; Vekemans et al., 2014), this could be due to divergence from the reference sequence. AL2G2623090 included sequences similar to both *Aly*10.1 (*ARK*1 in *A. thaliana*) and *Aly*10.2 (*ARK*2 in *A. thaliana*), which were detected in all individuals. *Aly*10.2 is a suspected pseudogene in *A. lyrata* due to a large deletion and does not amplify in all individuals (Charlesworth et al., 2003b), whereas *Aly*10.1 is predicted to be functional and amplifies in more individuals. Clustering suggested that only four individuals had both genes but not all contigs could be resolved due to missing parts of the sequence. AL6G484380 (*Aly*3) was found in all individuals. Fourteen of the contigs clustered into two distinct clades that did not show similarity to any known paralogs (contig-only clusters; **Supplementary Table 10**). One was found in 10/13 accessions while the other was only found in four; these could represent previously uncharacterized members of the *S*-receptor kinase gene family.

For the *SRK* sequences, five of the putatively new alleles found by 454 pyrosequencing were pulled out, all of which also were detected by cloning and sequencing using degenerate primers; multi-exon sequences were mined from the genomes for two of them (NEW2, NEW16; **Table 6**). Multi-exon sequences were also

TABLE 6 | Identity of *SRK*-like (AL7G32730) contigs pulled out by genome mining and confirmed by phylogenetic clustering.

Accession	Type	AL7G32730 (SRK)				Total linked	Unlinked				Total unlinked
SRR2040821	A4x	AlySRK01	AlySRK15	AlySRK42	**NEW17**	4					0
SRR2040822	A4x	AlySRK01 x 2	AHASRK08 x 2			4	Aly13-2*	AlySRK32			2
SRR2040825	A4x	AlySRK01	AlySRK12			2	**NEW2***				1
SRS945917	A2x	AHASRK17	**NEW16***			2	Aly13-2*	AlySRK32 x 2	AlySRK47		4
SRS1256176	A2x	AlySRK13				1	Aly13-2*	AlySRK32	AlySRK47	**NEW9**	4
SRS1256175	A2x	AlySRK01*	AlySRK23			2					0
SRR2040827	L4x	AlySRK01				1		AlySRK32	AlySRK47		2
SRR2040828	L4x	AlySRK01	AlySRK25	AlySRK33		3	Aly13-2	AlySRK32			2
SRR2040829[a]	L4x	AlySRK01	AlySRK12	**NEW17**	**NEW7**	6[a]			AlySRK47		1
SRR2040830	L4x	AlySRK01*	AlySRK42	AaSRK50		3					0
SRR3111440	L2x	AlySRK15*				1					0
SRR3111441	L2x	AlySRK44	AlySRK17			2	Aly13-7*	AlySRK32			2
SRR2040791	L2x	AlySRK01*	AlySRK42			2	**NEW2***				0

Contigs in red would have been mis-assigned based only on BLAST; those in blue were not resolved by BLAST. Cloned sequences were also obtained from contigs indicated in bold; asterisks indicate sequences where multi-exon sequences were pulled out using the genome mining. Alleles showing high similarity to SRK but not predicted to be linked to the SI phenotype are also indicted (Unlinked). The total number of linked and unlinked alleles resolved per accession is also indicated.

[a] *Also has AlySRK10 and AlySRK28.*

pulled out for *Aly*SRK01, *Aly*SRK15, *Aly*13-2, and *Aly*13-7. While the genome mining approach seems promising, the presence of homozygotes for *SRK* for three individuals suggests that not all *SRK* alleles were identified within individual genomes: one L2x and one A2x individual had a single dominant allele each (*AlySRK*15, in dominance Class A2 and *AlySRK*13 in dominance Class A3, respectively). One L4x individual was homozygous for *AlySRK*01, which is plausible, as homozygotes for this recessive allele have been found in previous segregation-based analyses of tetraploid *A. lyrata* from Austria (Mable et al., 2004).

Challenge: Extracting Full-Length Sequences of Polymorphic Genes From Short Read Data

While the genome mining holds promise for investigating copy number variation and obtaining full-length sequences from new alleles, the approach that worked best required a detailed reference database of alleles in order to accurately assign sequences to loci. BLAST analyses alone resulted in mis-assignment of *SRK* alleles to other paralogs and other paralogs were sometimes assigned as *SRK* alleles. While part of this was because not all sequences were available in Genbank for BLAST analysis, the gene conversion with unlinked loci that makes similarity alone unreliable (Prigoda et al., 2005) remained problematic in these analyses. For example, *Aly*13-2/13-7 sequences (which are not linked but are highly similar to Class B *SRK* alleles) were assigned as *SRK* in the initial analyses using only the five genes extracted from the MN47 genome. Manual alignments and phylogenetic clustering were required to determine allelic identities and to assign sequences to paralogous loci. However, there were clues in the BLAST analyses that suggested mis-assignment of *SRK*-like alleles; a signature of high E-value and low score in all cases predicted clustering to *SRK*-like sequences (although including unlinked loci such as *Aly*13-2). Nevertheless, the presence of only a single dominant allele in some accessions suggested that the genome mining did not pull out all *SRK* sequences that should have been present (since homozygotes should only be possible for recessive alleles).

While we had hoped also to be able to use this approach to map the potentially new alleles found using 454 sequencing to genomic regions to predict linkage to the *S*-locus, the failure of the mapping approach meant that this was not possible. However, when amplifying longer sequences using degenerate primers, we were able to obtain full-length sequences for some of the potentially new specificities predicted from the 454 analyses that we could use to BLAST the *de novo* assemblies (**Table 6**).

CONCLUSIONS AND RECOMMENDATIONS

The results presented here suggest that the highly polymorphic *SRK* alleles could be useful for interpreting evolutionary patterns of gene flow among populations, species and ploidy levels. We have demonstrated that tetraploids show no apparent advantage in terms of allelic or haplotypic repertoire due to more relaxed selection than diploids but that there is increased evidence for introgression (at least based on the most recessive *SRK* allele) among tetraploids from suspected hybrid populations. We also demonstrated that following up high throughput genotyping with targeted PCR can help to increase accuracy and completeness. We also identified new alleles not previously characterized and predicted dominance based on phylogenetic clustering.

Nevertheless, there are some important caveats from the analyses, which highlight considerations for future studies based on more robust approaches to high throughput genotyping. We make the following recommendations for future investigations of gene family evolution, in diploids as well as polyploids: (1) applying a hierarchical strategy to filtering decisions for cluster analyses could improve assignment of sequence variants to allelic variants, similar to suggestions for hierarchical AMOVA or STRUCTURE analyses (Holsinger and Mason-Gamer, 1996;

Herdegen et al., 2014); (2) amplicon-based approaches for genotyping using deep sequencing should be avoided if there are other options available, as differential amplification and the difficulty of distinguishing PCR errors from real biological processes are difficult to overcome by any current sequencing technology; (3) due to the difficulty of assigning variants to gene copies, interpretation of gene family evolution should always be accompanied by co-segregation of sequence variants with the phenotype, whenever possible; (4) genome mining of resequenced genomes has the potential to investigate copy number variation and obtain full-length sequences that would be useful for population genetics analyses and tests for selection but lack of assembly of highly polymorphic genes to references means that this might only be practical for genes where there is already extensive knowledge about the components of the gene family.

While our results have demonstrated some useful insights into the dynamics of a complex gene family in polyploids and hybrids, we recommend that non-PCR-based sequence capture approaches hold the most promise for assessing patterns of selection on genes under balancing selection, where trans-specific polymorphism, reduced differentiation among alleles, and intermediate frequency alleles are predicted. Such approaches, for example, have been successfully applied to investigating R-gene variation in crop plants (Jupe et al., 2012, 2013; Andolfo et al., 2014; Giolai et al., 2016; Russell et al., 2016; Van Weymers et al., 2016). Whole genome resequencing approaches could be useful for setting the genomic context and fate of duplications, but there are still substantial challenges to resolve in distinguishing loss of copies from lack of coverage or lack of assembly to the reference due to high sequence divergence. A hierarchical approach to filtering or assembly to multiple references (e.g., multiple individuals or multiple alleles or gene family members) could help to overcome such difficulties but resolving fine-scale variation among variants from errors (e.g., haplotypes within specificities) and resolving complete heterozygous genotypes (particularly in polyploids) will require some creative bioinformatic solutions.

DATA AVAILABILITY STATEMENT

The 200 bp fragments generated by 454 sequencing are too short for submission to Genbank but a full alignment of the sequences identified has been provided as **Supplementary Data Sheet 1** (including only the 454 sequences and references), 2 (including all unique alleles found across analyses), and 3 (*SRK*01 sequences). All new Sanger sequences have been deposited to Genbank (Accession numbers: MH507371-MH507400). Accession numbers and details for all unique sequences identified, along with those for reference sequences already available in Genbank are provided in **Supplementary Table 12**.

AUTHOR CONTRIBUTIONS

BM wrote the paper and performed the bulk of the sequence analyses described. MJ, AC, and PR-D generated sequence data for objectives 1, 2, and 3, respectively. MJ also performed the crosses for segregation analyses. KL and CK developed bespoke bioinformatics pipelines for objectives 1 and 4, respectively. AB contributed conceptually and financially to objectives 1 and 2. MK contributed conceptually to all aspects of the project and provided samples and advice on sampling locations and interpretation of the hybrid zones. AB, MK, and CK contributed to writing the manuscript.

FUNDING

This project was funded by a NERC Advanced Research Fellowship (NE/B50094X/1) to BM, and support from the Centre for Ecology and Evolutionary Synthesis (RCN/179569) to AB. Further support for genome resequencing was granted through a DFG grant (DFG SPP 1529; KO2302-14) to MK.

ACKNOWLEDGMENTS

We thank Aileen Adam and Elizabeth Kilbride for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2018.00114/full#supplementary-material

REFERENCES

Adams, K. L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered.* 98, 136–141. doi: 10.1093/jhered/esl061

Aeschlimann, P. B., Häberli, M. A., Reusch, T. B. H., Boehm, T., and Milinski, M. (2003). Female sticklebacks *Gasterosteus aculeatus* use self-reference to optimize MHC allele number during mate selection. *Behav. Ecol. Sociobiol.* 54, 119–126. doi: 10.1007/s00265-003-0611-6

Andolfo, G., Jupe, F., Witek, K., Etherington, G. J., Ercolano, M. R., and Jones, J. D. G. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biol.* 14:120. doi: 10.1186/1471-2229-14-120

Arnold, B., Kim, S. T., and Bomblies, K. (2015). Single geographic origin of a widespread autotetraploid arabidopsis arenosa lineage followed by interploidy admixture. *Mol. Biol. Evol.* 32, 1382–1395. doi: 10.1093/molbev/msv089

Bandelt, H. J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Bechsgaard, J., Hedegaard Jorgensen, T., and Schierup, M. (2017). Evidence for adaptive introgression of disease resistance genes among closely related arabidopsis species. G3 7, 2677–2683. doi: 10.1534/g3.117.043984

Bechsgaard, J. S., Castric, V., Charlesworth, D., Vekemans, X., and Schierup, M. H. (2006). The transition to self-compatibility in *Arabidopsis thaliana* and evolution within *S*-haplotypes over 10 Myr. *Mol. Biol. Evol.* 23, 1741–1750. doi: 10.1093/molbev/msl042

Beckmann, J. S., Estivill, X., and Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet*. 8, 639–646. doi: 10.1038/nrg2149

Billiard, S., Castric, V., and Vekemans, X. (2007). A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. *Genetics* 175, 1351–1369. doi: 10.1534/genetics.105.055095

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New. Phyt*. 186, 54–62. doi: 10.1111/j.1469-8137.2009.03087.x

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi: 10.1105/tpc.021345

Boggs, N. A., Dwyer, K. G., Shah, P., McCuuoch, A. A., Bechsgaard, J., Schierup, M. H. et al. (2009). Expression of distinct self-incompatibility specificities in *Arabidopsis thaliana*. *Genetics* 182, 1313–1321. doi: 10.1534/genetics.109.102442

Busch, J. W., Sharma, J., and Schoen, D. J. (2008). Molecular characterization of *Lal*2, an *SRK*-like gene linked to the S-locus in the wild mustard *Leavenworthia alabamica*. *Genetics* 178, 2055–2067. doi: 10.1534/genetics.107.083204

Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T., and Kauserud, H. (2012). Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol.* 5, 747–749. doi: 10.1016/j.funeco.2012.06.003

Castric, V., Bechsgaard, J., Schierup, M. H., and Vekemans, X. (2008). Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet*. 4:e1000168. doi: 10.1371/journal.pgen.1000168

Castric, V., Bechsgaard, J. S., Grenier, S., Noureddine, R., Schierup, M. H., and Vekemans, X. (2010). Molecular evolution within and between self-incompatibility specificities. *Mol. Biol. Evol*. 27, 11–20. doi: 10.1093/molbev/msp224

Castric, V., and Vekemans, X. (2007). Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? *BMC Evol. Biol*. 7:132. doi: 10.1186/1471-2148-7-132

Charlesworth, D., Awadalla, P., Mable, B. K., and Schierup, M. H. (2000). Population-level studies of multiallelic self-incompatibility loci, with particular reference to Brassicaceae. *Ann. Bot*. 85, 227–239. doi: 10.1006/anbo.1999.1015

Charlesworth, D., Bartolome, C., Schierup, M. H., and Mable, B. K. (2003a). Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. *Mol. Biol. Evol*. 20, 1741–1753. doi: 10.1093/molbev/msg170

Charlesworth, D., Kamau, E., Hagenblad, J., and Tang, C. L. (2006). Trans-specificity at loci near the self-incompatibility loci in *Arabidopsis*. *Genetics* 172, 2699–2704. doi: 10.1534/genetics.105.051938

Charlesworth, D., Mable, B. K., Schierup, M. H., Bartolome, C., and Awadalla, P. (2003b). Diversity and linkage of genes in the self-incompatibility gene family in *Arabidopsis lyrata*. *Genetics* 164, 1519–1535.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., et al. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar. Genomics* 30, 3–13. doi: 10.1016/j.margen.2016.04.012

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., et al. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Gen*. 17:55. doi: 10.1186/s12864-015-2194-9

Dart, S., Kron, P., and Mable, B. K. (2004). Characterizing polyploidy in *Arabidopsis lyrata* using chromosome counts and flow cytometry. *Can. J. Bot*. 82, 185–197. doi: 10.1139/b03-134

Delport, W., Poon, A. F., Frost, S. D., and Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 2455–2457. doi: 10.1093/bioinformatics/btq429

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem.Bull*. 19, 11–15.

Duvaux, L., Geissmann, Q., Gharbi, K., Zhou, J. J., Ferrari, J., Smadja, C. M., et al. (2015). Dynamics of copy number variation in host races of the pea aphid. *Mol. Biol. Evol*. 32, 63–80. doi: 10.1093/molbev/msu266

Evans, B. J. (2007). Ancestry influences the fate of duplicated genes millions of years after polyploidization of clawed frogs (*Xenopus*). *Genetics* 176, 1119–1130. doi: 10.1534/genetics.106.069690

Flajnik, M. F., and Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet*. 11, 47–59. doi: 10.1038/nrg2703

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545.

Foxe, J. P., Stift, M., Tedder, A., Haudry, A., Wright, S. I., and Mable, B. K. (2010). Reconstructing origins of loss of self-incompatibility and selfing in North American *Arabidopsis lyrata*: a population genetic context. *Evolution* 64, 3495–3510. doi: 10.1111/j.1558-5646.2010.01094.x

Gardiner, L. J., Bansept-Basler, P., Olohan, L., Joynson, R., Brenchley, R., Hall, N., et al. (2016). Mapping-by-sequencing in complex polyploid genomes using genic sequence capture: a case study to map yellow rust resistance in hexaploid wheat. *Plant J*. 87, 403–419. doi: 10.1111/tpj.13204

Giolai, M., Paajanen, P., Verweij, W., Percival-Alwyn, L., Baker, D., Witek, K., et al. (2016). Targeted capture and sequencing of gene-sized DNA molecules. *Biotechniques* 61, 315–322. doi: 10.2144/000114484

Goubet, P. M., Berges, H., Bellec, A., Prat, E., Helmstetter, N., Mangenot, S., et al. (2012). Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS Genet*. 8:e1002495. doi: 10.1371/journal.pgen.1002495

Gout, J. F., and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol. Biol. Evol*. 32, 2141–2148. doi: 10.1093/molbev/msv095

Guggisberg, A., Mansion, G., and Conti, E. (2009). Disentangling reticulate evolution in an arctic-alpine polyploid complex. *Syst. Biol*. 58, 55–73. doi: 10.1093/sysbio/syp010

Guo, Y. L., Zhao, X., Lanz, C., and Weigel, D. (2011). Evolution of the S-locus region in *Arabidopsis* relatives. *Plant Physiol*. 157, 937–946. doi: 10.1104/pp.111.174912

Hatakeyama, K., Watanabe, M., Takasaki, T., Ojima, K., and Hinata, K. (1998). Dominance relationships between *S*-alleles in self-incompatible *Brassica campestris* L. *Heredity* 80, 241–247. doi: 10.1046/j.1365-2540.1998.00295.x

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res*. 44, W236–W241. doi: 10.1093/nar/gkw370

Herdegen, M., Babik, W., and Radwan, J. (2014). Selective pressures on MHC class II genes in the guppy (*Poecilia reticulata*) as inferred by hierarchical analysis of population structure. *J. Evol. Biol*. 27, 2347–2359. doi: 10.1111/jeb.12476

Hermansen, R. A., Hvidsten, T. R., Sandve, S. R., and Liberles, D. A. (2016). Extracting functional trends from whole genome duplication events using comparative genomics. *Biol. Proc. Online* 18:11. doi: 10.1186/s12575-016-0041-2

Hohmann, N., and Koch, M. A. (2017). An *Arabidopsis* introgression zone studied at high spatio-temporal resolution: interglacial and multiple genetic contact exemplified using whole nuclear and plastid genomes. *BMC Gen*. 18:6. doi: 10.1186/s12864-017-4220-6

Hohmann, N., Schmickl, R., Chiang, T. Y., Lu Anova, M., Kola, F., Marhold, K., et al. (2014). Taming the wild: resolving the gene pools of non-model *Arabidopsis lineages*. *BMC Evol. Biol*. 14:224. doi: 10.1186/s12862-014-0224-x

Holsinger, K., and Mason-Gamer, R. J. (1996). Hierarchical analysis of nucleotide diversity in geographically structured populations. *Genetics* 142, 629–639.

Hull, R. M., Cruz, C., Jack, C. V., and Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. *PLoS Biol*. 15:e2001333. doi: 10.1371/journal.pbio.2001333

Jørgensen, M. H., Ehrich, D., Schmickl, R., Koch, M. A., and Brysting, A. K. (2011). Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae). *BMC Evol. Biol*. 11:346. doi: 10.1186/1471-2148-11-346

Jørgensen, M. H., Lagesen, K., Mable, B. K., and Brysting, A. K. (2012). Using high-throughput sequencing to investigate the evolution of self-incompatibility genes in the Brassicaceae: strategies and challenges. *Plant Ecol. Divers* 5, 473–484. doi: 10.1080/17550874.2012.748098

Jupe, F., Pritchard, L., Etherington, G. J., Mackenzie, K., Cock, P. J., and Wright, F. et al. (2012). Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Gen*. 13:75. doi: 10.1186/1471-2164-13-75

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G. J. et al. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* 76, 530–544. doi: 10.1111/tpj.12307

Kalbe, M., Eizaguirre, C., Dankert, I., Reusch, T. B. H., Sommerfeld, R. D., and Wegner, K. M. (2009). Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proc. R. Soc. B*. 276, 925–934. doi: 10.1098/rspb.2008.1466

Koch, M. A., Wernisch, M., and Schmickl, R. (2008). *Arabidopsis thaliana*'s wild relatives: an updated overview on systematics, taxonomy and evolution. *Taxon* 57, 933–943.

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., et al. (2017). Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 114, E913–E921. doi: 10.1073/pnas.1619268114

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

Kusaba, M., Dwyer, K., Hendershot, J., Vrebalov, J., Nasrallah, J. B., and Nasrallah, M. E. (2001). Self-incompatibility in the genus *Arabidopsis*: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana. Plant Cell* 13, 627–643. doi: 10.1105/tpc.13.3.627

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Leducq, J. B., Gosset, C. C., Gries, R., Calin, K., Schmitt, E., Castric, V., et al. (2014). Self-incompatibility in Brassicaceae: identification and characterization of SRK-like sequences linked to the S-locus in the tribe Biscutelleae. *Genes Genom. Genet*. 4, 983–992. doi: 10.1534/g3.114.010843

Leigh, J. W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Meth. Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Lewis, D. (1947). Competition and dominance of incompatibility alleles in diploid pollen. *Heredity* 1, 85–108. doi: 10.1038/hdy.1947.5

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

Llaurens, V., Billiard, S., Castric, V., and Vekemans, X. (2009). Evolution of dominance in sporophytic self-incompatibility systems: I. Genetic load and coevolution of levels of dominance in pollen and pistil. *Evolution* 63, 2427–2437. doi: 10.1111/j.1558-5646.2009.00709.x

Llaurens, V., Billiard, S., Leducq, J. B., Castric, V., Klein, E. K., and Vekemans, X. (2008). Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* 62, 2545–2557. doi: 10.1111/j.1558-5646.2008.00469.x

Luikart, G., Allendorf, F. W., Cornuet, J. M., and Sherwin, W. B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.* 89, 238–247. doi: 10.1093/jhered/89.3.238

Mable, B. K. (2013). Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? *Heredity* 110, 95–96. doi: 10.1038/hdy.2012.105

Mable, B. K., and Adam, A. (2007). Patterns of genetic diversity in outcrossing and selfing populations of *Arabidopsis lyrata*. *Mol. Ecol*. 16, 3565–3580. doi: 10.1111/j.1365-294X.2007.03416.x

Mable, B. K., Beland, J., and Di Berardo, C. (2004). Inheritance and dominance of self-incompatibility alleles in polyploid *Arabidopsis lyrata*. *Heredity* 93, 476–486. doi: 10.1038/sj.hdy.6800526

Mable, B. K., Hagmann, J., Kim, S. T., Adam, A., Kilbride, E., Weigel, D., et al. (2017). What causes mating system shifts in plants? *Arabidopsis lyrata* as a case study. *Heredity* 118, 52–63. doi: 10.1038/hdy.2016.99

Mable, B. K., Kilbride, E., Viney, M. E., and Tinsley, R. C. (2015). Copy number variation and genetic diversity of MHC Class IIb alleles in an alien population of *Xenopus laevis*. *Immunogenetics* 67, 591–603. doi: 10.1007/s00251-015-0860-3

Mable, B. K., Schierup, M. H., and Charlesworth, D. (2003). Estimating the number, frequency, and dominance of S-alleles in a natural population of *Arabidopsis lyrata* (Brassicaceae) with sporophytic control of self-incompatibility. *Heredity* 90, 422–431. doi: 10.1038/sj.hdy.6800261

Maddison, D. R., and Maddison, W. P. (2000). *Macclade 4: Analysis Of Phylogeny And Character Evolution, Version 4.0 edn*. Sunderland, MA: Sinauer Associates.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011:17. doi: 10.14806/ej.17.1.200

Mattenberger, F., Sabater-Munoz, B., Toft, C., Sablok, G., and Fares, M. A. (2017). Expression properties exhibit correlated patterns with the fate of duplicated genes, their divergence, and transcriptional plasticity in Saccharomycotina. *DNA Res* 24, 559–570. doi: 10.1093/dnares/dsx025

Meyer, A., and Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27, 937–945. doi: 10.1002/bies.20293

Muir, G., Ruiz-Duarte, P., Hohmann, N., Mable, B. K., Novikova, P., Schmickl, R., et al. (2015). Exogenous selection rather than cytonuclear incompatibilities shapes asymmetrical fitness of reciprocal *Arabidopsis* hybrids. *Ecol. Evol*. 5, 1734–1745. doi: 10.1002/ece3.1474

Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., et al. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet*. 48, 1077–1082. doi: 10.1038/ng.3617

Ohno, S. (1970). *Evolution by Gene Duplication*. New York, NY: Springer-Verlag.

Paetsch, M., Mayland-Quellhorst, S., and Neuffer, B. (2006). Evolution of the self-incompatibility system in the Brassicaceae: identification of *S*-locus receptor kinase (SRK) in self-incompatible *Capsella grandiflora*. *Heredity* 97, 283–290. doi: 10.1038/sj.hdy.6800854

Pond, K. S. L., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679. doi: 10.1093/bioinformatics/bti079

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. W. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098. doi: 10.1093/bioinformatics/btl474

Prigoda, N. L., Nassuth, A., and Mable, B., K (2005). Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. *Mol. Biol. Evol*. 22, 1609–1620. doi: 10.1093/molbev/msi153

Rambaut, A. (1996). *Se-Al: Sequence Alignment Editor, Version 1.0 Alpha1*. Distributed Over the World Wide Web. Available online at: http://tree.bio.ed.ac.uk/software/seal/

Reusch, T. B. H., Haberli, M. A., Aeschlimann, P. B., and Milinski, M. (2001). Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. *Nature* 414, 300–302. doi: 10.1038/35104547

Rodrigo, G., and Fares, M. A. (2018). Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. *Elife* 7:29739. doi: 10.7554/eLife.29739

Ruiz-Duarte, P. (2012). *Self Incompatibility Alleles In Wild Relatives Of Arabidopsis Thaliana*. PhD thesis, University of Heidelberg, Heidelberg.

Russell, J., Mascher, M., Dawson, I. K., Kyriakidis, S., Calixto, C., Freund, F., et al. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet*. 48, 1024–1030. doi: 10.1038/ng.3612

Saintenac, C., Jiang, D. Y., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol*. 12:r88. doi: 10.1186/gb-2011-12-9-r88

Salmon, A., Udall, J. A., Jeddeloh, J. A., and Wendel, J. (2012). Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. *Genes Genom. Genet*. 2, 921–930. doi: 10.1534/g3.112.003392

Schierup, M. H., Mable, B. K., Awadalla, P., and Charlesworth, D. (2001). Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. *Genetics* 158, 387–399.

Schierup, M. H., Vekemans, X., and Christiansen, F. B. (1998). Allelic geneologies in sporophytic self-incompatibility systems in plants. *Genetics* 150, 1187–1198.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17:976. doi: 10.1186/s12859-016-0976-y

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 43:1341. doi: 10.1093/nar/gku1341

Schmickl, R. E. (2009). *Reticulate Evolution in Glacial Refuge Areas-the Genus Arabidopsis in the Eastern Austrian Danube Valley (Wachau)*. PhD thesis, Ruperto-Carolo University of Heidelberg.

Schmickl, R., Jørgensen, M. H., Brysting, A. K., and Koch, M. A. (2010). The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol. Biol*. 10:98. doi: 10.1186/1471-2148-10-98

Schmickl, R., and Koch, M. A. (2011). *Arabidopsis* hybrid speciation processes. *Proc. Natl. Acad. Sci. U.S.A*. 108, 14192–14197. doi: 10.1073/pnas.1104212108

Schoen, D. J., and Busch, J. W. (2009). The evolution of dominance in sporophytic self-incompatibility systems. II. Mate availability and recombination. *Evolution* 63, 2099–2113. doi: 10.1111/j.1558-5646.2009.00686.x

Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29, 2790–2791. doi: 10.1093/bioinformatics/btt468

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour*. 11, 1–8. doi: 10.1111/j.1755-0998.2010.02979.x

Seoighe, C., and Gehring, C. (2004). Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet*. 20, 461–464. doi: 10.1016/j.tig.2004.07.008

Shiba, H., Iwano, M., Entani, T., Ishimoto, K., Shimosato, H., Che, F. S., et al. (2002). The dominance of alleles controlling self-incompatibility in Brassica pollen in regulated at the RNA level. *Plant Cell* 14, 491–504. doi: 10.1105/tpc.010378

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., and Li, W., et al (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Sys. Biol*. 7:539. doi: 10.1038/msb.2011.75

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., and Kent, M. P., Nome, T. et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. doi: 10.1038/nature17164

Soltis, P. S., Doyle, J., and Soltis, D. E. (eds.). (1992). "Molecular data and polyploid evolution in plants," in *Molecular Systematics of Plants* (New York, NY: Chapman and Hall), 177–201.

Stevens, J. P., and Kay, Q. O. N. (1989). The number, dominance relationships and frequencies of self-incompatibility alleles in a natural population of *Sinapis arvensis* L. in South Wales. *Heredity* 62, 199–205. doi: 10.1038/hdy.1989.29

Stoeckel, S., Castric, V., Mariette, S., and Vekemans, X. (2008). Unequal allelic frequencies at the self-incompatibility locus within local populations of *Prunus avium* L.: an effect of population structure? *J. Evol. Biol*. 21, 889–900. doi: 10.1111/j.1420-9101.2008.01504.x

Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-specific evolution of polymorphism. *Proc. Natl. Acad. Sci. U.S.A*. 87, 2419–2423.

Tedder, A., Ansell, S. W., Lao, X., Vogel, J. C., and Mable, B. K. (2011). Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*. *Ann. Bot*. 108, 699–713. doi: 10.1093/aob/mcr157

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

van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., et al. (2007). Complexity reduction of polymorphic sequences (CRoPS^{TM}): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2:e1172. doi: 10.1371/journal.pone.0001172

Van Weymers, P. S. M., Baker, K., Chen, X., Harrower, B., Cooke, D. E. L., Gilroy, E. M., et al. (2016). Utilizing "omic" yechnologies to identify and prioritize novel sources of resistance to the oomycete pathogen *Phytophthora infestans* in potato germplasm collections. *Front. Plant Sci*. 7:672. doi: 10.3389/fpls.2016.00672

Vekemans, X., Leducq, J. B., Llaurens, V., Castric, V., Saumitou-Laprade, P., and Hardy, O. J. (2011). Effect of balancing selection on spatial genetic structure within populations: theoretical investigations on the self-incompatibility locus and empirical studies in *Arabidopsis halleri*. *Heredity* 106, 319–329. doi: 10.1038/hdy.2010.68

Vekemans, X., Poux, C., Goubet, P. M., and Castric, V. (2014). The evolution of selfing from outcrossing ancestors in Brassicaceae: what have we learned from variation at the S-locus? *J. Evol. Biol*. 27, 1372–1385. doi: 10.1111/jeb.12372

Wegner, K. M., Kalbe, M., Kurtz, J., Reusch, T. B. H., and Milinski, M. (2003). Parasite selection for immunogenic optimality. *Science* 301:1343. doi: 10.1126/science.1088293

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., and McGuire, A., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–U875. doi: 10.1038/nature06884

Wolfe, K. H., and Shields, D., C (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713. doi: 10.1038/42711

Xing, S. L., Li, M. Y., and Liu, P. (2013). Evolution of S-domain receptor-like kinases in land plants and origination of S-locus receptor kinases in Brassicaceae. *BMC Evol. Biol*. 13:69. doi: 10.1186/1471-2148-13-69

Zmienko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor. Appl. Genet*. 127, 1–18. doi: 10.1007/s00122-013-2177-7

Conflict of Interest Statement: 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 23 May 2018
doi: 10.3389/fevo.2018.00066

Analysis of Molecular Variance (AMOVA) for Autopolyploids

Patrick G. Meirmans[1]* and *Shenglin Liu*[2]

[1] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands, [2] Department of Bioscience, Aarhus University, Aarhus, Denmark

Autopolyploids present several challenges to researchers studying population genetics, since almost all population genetics theory, and the expectations derived from this theory, has been developed for haploids and diploids. Also many statistical tools for the analysis of genetic data, such as AMOVA and genome scans, are available only for haploids and diploids. In this paper, we show how the Analysis of Molecular Variance (AMOVA) framework can be extended to include autopolyploid data, which will allow calculating several genetic summary statistics for estimating the strength of genetic differentiation among autopolyploid populations (F_{ST}, φ_{ST}, or R_{ST}). We show how this can be done by adjusting the equations for calculating the Sums of Squares, degrees of freedom and covariance components. The method can be applied to a dataset containing a single ploidy level, but also to datasets with a mixture of ploidy levels. In addition, we show how AMOVA can be used to estimate the summary statistic ρ, which was developed especially for polyploid data, but unfortunately has seen very little use. The ρ-statistic can be calculated in an AMOVA by first calculating a matrix of squared Euclidean distances for all pairs of individuals, based on the within-individual allele frequencies. The ρ-statistic is well suited for polyploid data since its expected value is independent of the ploidy level, the rate of double reduction, the frequency of polysomic inheritance, and the mating system. We tested the method using data simulated under a hierarchical island model: the results of the analyses of the simulated data closely matched the values derived from theoretical expectations. The problem of missing dosage information cannot be taken into account directly into the analysis, but can be remedied effectively by imputation of the allele frequencies. We hope that the development of AMOVA for autopolyploids will help to narrow the gap in availability of statistical tools for diploids and polyploids. We also hope that this research will increase the adoption of the ploidy-independent ρ-statistic, which has many qualities that makes it better suited for comparisons among species than the standard F_{ST}, both for diploids and for polyploids.

Keywords: genetic differentiation, population structure, F_{ST}, double reduction, polysomic inheritance, polyploidy, AMOVA

OPEN ACCESS

Edited by:
Richard John Abbott,
University of St. Andrews,
United Kingdom

Reviewed by:
Lindsay V. Clark,
University of Illinois at
Urbana-Champaign, United States
Peter E. Smouse,
Rutgers University, The State
University of New Jersey,
United States

****Correspondence:***
Patrick G. Meirmans
p.g.meirmans@uva.nl

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

Received: *01 February 2018*
Accepted: *30 April 2018*
Published: *23 May 2018*

Citation:
Meirmans PG and Liu S (2018)
Analysis of Molecular Variance
(AMOVA) for Autopolyploids.
Front. Ecol. Evol. 6:66.
doi: 10.3389/fevo.2018.00066

INTRODUCTION

Autopolyploidy is an important, but often overlooked, aspect of the evolution of all major groups of Eukaryotes-plants, animals, and fungi- and may constitute an underappreciated source of biodiversity (Hardy, 2015). There are many species in which multiple ploidy levels (cytotypes) exist and often each cytotype itself conforms to the requirements of several widely used species concepts (Soltis et al., 2007). Autopolyploidy has many effects on the mechanisms of evolution, not only because of the increase in genomic content and the flexibility for developing new traits (Larkin et al., 2016), but also because, compared to diploidy, it generates different dynamics of allele frequencies that interact with various demographic processes, influencing adaptation and speciation (Parisod et al., 2010). In a species with different ploidy levels, the different cytotypes often show intricate geographical patterns in their distribution, which may be the result of historical, demographic, ecological, or genetic processes (Glennon et al., 2014; Kolár et al., 2017). The analysis of population genetic structure of autopolyploids may therefore reveal a lot about these processes. However, polyploids also present several challenges to the researchers studying their population genetics (Dufresne et al., 2014). This is because population genetic theory, the expectations derived from this theory, and the statistical tools for data analysis were developed mostly for haploids and diploids and require translation for polyploids (Meirmans et al., 2018).

Several of the basic genetic processes work differently in autopolyploids than in diploids (Meirmans et al., 2018). The higher number of chromosomes means that for each gene a higher total number of copies is present in a population. This increases the number of mutation events per population, and also increases the impact of migration as each migrant individual carries more chromosome copies to its new population. Conversely, the higher total number of chromosome copies is akin to a higher effective population size and therefore reduces the force of genetic drift, compared to a diploid population with the same number of individuals. Mendelian segregation also works differently in autopolyploids, since it is not necessarily completely random, as is almost always the case in diploids. Instead, there may be disomic inheritance, polysomic inheritance, or a combination of the two, where the rate of polysomy varies across the genome (Stift et al., 2008; Meirmans and van Tienderen, 2013). In addition, autopolyploids may show double reduction, a process where two copies of the same chromatid segment end up in the same gamete (Bever and Felber, 1992; Hardy, 2015). For example in a autotetraploid with genotype *ABCD* this may lead to the production of homozygous (*AA*, *BB*, *CC*, and *DD*) gametes, in addition to the expected heterozygous gametes (e.g., *AB*, *AD*). A more practical problem in the genetic analysis of polyploids is that it is often difficult to estimate the dosage of the different alleles in a genotype (Dufresne et al., 2014). For example, it may be impossible to distinguish between the triploid genotypes *AAB* and *ABB* since they both share the marker phenotype *AB*. Missing dosage information may introduce a bias in the subsequent analysis; though depending on the type of analysis this bias may be corrected for quite effectively when random mating in populations can be assumed (De Silva et al., 2005; Meirmans et al., 2018). However, when the assumption of Hardy Weinberg equilibrium cannot be made for a species, accounting for the missing dosage information becomes more problematic, though in some cases it is possible to adjust the calculations specifically to take the missing dosage into account (Hardy, 2015; Field et al., 2017).

Estimating the strength of the genetic population structure is usually done using *F*-statistics that decompose the genetic variance into within-individual, within-population and among-population components (Wright, 1969). Autopolyploidy affects the way these statistics should be estimated (Meirmans et al., 2018), but also their expected values under a given model of population structure, when compared to the same model for diploids (Ronfort et al., 1998). For example, the expected value of F_{ST}—quantifying the degree of population differentiation—depends on the balance among migration, mutation, and drift. In autopolyploids, the increased effects of mutation and migration, in combination with the reduced force of drift, cause the expected value of F_{ST} to be lower than the corresponding value for diploids (Meirmans et al., 2018). This difference in expectation complicates comparisons of the strength of population structure among species or sets of populations with different ploidy levels.

To enable a better estimation of the degree of population differentiation across ploidy levels, Ronfort et al. (1998) developed an alternative summary statistic, which they called ρ, for which the expected value is independent of the ploidy level. The ρ-statistic is comparable to F_{ST} in that it estimates the degree of population differentiation and—barring estimation error—ranges between 0 and 1. For haploid data, the value of ρ is exactly the same as the value of F_{ST}; for higher ploidy levels, the value of ρ is generally slightly higher than that of F_{ST}. The ploidy independence of ρ is achieved by disregarding the within-individual variation (illustrated by Equation 14 below). Another perk of the ρ-statistic that makes it suitable for the analysis of polyploid data is that its value is both independent of the rate of double reduction (Ronfort et al., 1998) and of the frequency of polysomic inheritance (Meirmans and van Tienderen, 2013). The ρ-statistic also has a major advantage that is applicable to diploid as well as polyploid data: its value is independent of the rate of self-fertilization or other forms of inbreeding. This means that under a given model of population structure, ρ will have the same value for a strict inbreeder as for an obligate outcrosser, whereas F_{ST} gives higher values for inbreeders than for outcrossers. This is especially useful in comparative studies, where a comparison of F_{ST} and ρ can be used to see whether differences in population structure are due to differences in mating system or due to differences in population connectivity. Unfortunately, ρ is not very widely used, possibly because there are only few computer programs that allow estimation of ρ from genetic marker data. The only two such programs that we are aware of are SpaGedi (Hardy and Vekemans, 2002), and GenoDive (Meirmans and van Tienderen, 2004).

One of the most popular methods for estimating *F*-statistics is via Analysis of Molecular Variance (AMOVA) (Excoffier et al., 1992; Peakall et al., 1995; Michalakis and Excoffier, 1996). This popularity is probably due to the remarkable flexibility of the AMOVA framework: it can be used for the estimation of different types of *F*-statistics (F_{ST}, φ_{ST}, R_{ST}) and can easily incorporate additional hierarchical levels of population structure (e.g., testing for differentiation among groups of populations). In addition, AMOVA can be used to detect population clustering in a genetic dataset (Dupanloup et al., 2002; Meirmans, 2012). However, AMOVA has been described only for haploid and diploid data and the link between the AMOVA framework and the ρ-statistic has not been explored theoretically.

In this paper, we outline how the AMOVA framework can be extended to include autopolyploid data. We start by discussing how the standard AMOVA, for calculating F_{ST}, φ_{ST}, or R_{ST}, can be easily adapted for use with autopolyploids. We then show how the ploidy-independent ρ-statistic can be calculated in AMOVA by using a matrix of squared Euclidean distances between individuals, calculated from the within-individual allele frequencies. Finally, we show the application of the method by calculating both F_{ST} and ρ for simulated datasets and discuss how to deal with the polyploidy-specific complication of missing dosage information.

THE AMOVA FRAMEWORK

General Approach

In AMOVA (Excoffier et al., 1992; Michalakis and Excoffier, 1996), F-statistics are calculated from a set of covariance components, corresponding to the different hierarchical levels assumed to be present in the population structure (following Cockerham, 1973; Weir and Cockerham, 1984). So under a simple model of population structure where individuals are distributed over a number of populations, we can decompose the total genetic variance (σ_T^2) into among-populations (σ_a^2), among-individuals within populations (σ_b^2), and within-individuals (σ_c^2) covariance components, such that $\sigma_T^2 = \sigma_a^2{+}\sigma_b^2{+}\sigma_c^2$. The F-statistics can then be calculated as simple ratios of those covariance components:

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2} \tag{1a}$$

$$F_{IS} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2} \tag{1b}$$

$$F_{IT} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} \tag{1c}$$

When the populations can be clustered into multiple groups, an extra hierarchical level is added and the total genetic variance is decomposed into among-groups (σ_a^2), among-populations within-groups (σ_b^2), among-individuals within-populations (σ_c^2), and within individuals (σ_d^2) covariance components, such that $\sigma_T^2 = \sigma_a^2{+}\sigma_b^2{+}\sigma_c^2{+}\sigma_d^2$. The corresponding F-statistics are then:

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2} \tag{2a}$$

$$F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2 + \sigma_d^2} \tag{2b}$$

$$F_{IS} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_d^2} \tag{2c}$$

$$F_{IT} = \frac{\sigma_a^2 + \sigma_b^2 + \sigma_c^2}{\sigma_T^2} \tag{2d}$$

This follows the Analysis of Variance framework that was developed earlier by Cockerham (1973) and Weir and Cockerham (1984). However, whereas Weir and Cockerham calculated these covariance components from a linear vector of allele frequencies, AMOVA calculates them using a matrix $\boldsymbol{D}$ of pairwise squared Euclidean distances. This is based on previous work by Li (1976) showing that conventional Sums of Squares can be calculated from a matrix of pairwise squared Euclidean distances. These Sums of Squares can then be used to calculate the Expected Mean Squares, which in turn can be used to calculate the covariance components (Weir and Cockerham, 1984).

The use of a distance metric is actually what gives AMOVA its remarkable flexibility, as the distance metric can be changed, depending on the type of data under analysis. A simple matching distance can be used for a single locus with allelic data—for example for SNPs (Peakall et al., 1995; Michalakis and Excoffier, 1996). Multilocus values of the F-statistics can then be obtained by summing the covariance components over loci. A distance metric for haplotypic data was described in the original paper by Excoffier et al. (1992), based on the phenetic distance between the pair of haplotypes. This is also the most frequently used method for sequence data, though more complex distance metrics can be used as well—e.g., by incorporating a specific mutational model or by tracing distances along a connecting network or tree (Excoffier and Smouse, 1994). A distance metric for microsatellites loci can be calculated by taking the squared difference in repeat number between alleles (Michalakis and Excoffier, 1996).

The interpretation of the F-statistics returned by AMOVA depends strongly on the choice of distance metric used. This means that from the wide array of available estimators for F_{ST}, different estimators are obtained by different distance metrics. For allelic data, where the simple matching distance is used, the resulting F-statistics are mathematically equivalent to the estimators of Weir and Cockerham (1984). In contrast, for haplotypic/sequence data, the distances are indicative of the evolutionary relationships between haplotypes/sequences (Whitlock, 2011); to reflect this, the F-statistics are generally referred to with the Greek letter φ. Finally, when for microsatellites the difference in repeat number is used, the estimator corresponds to the R_{ST}-statistic (Slatkin, 1995).

Adaptation to Autopolyploids

For autopolyploids, AMOVA can be performed using the same methods as above for calculating the pairwise distances among alleles, yielding estimates of F_{ST}, φ_{ST}, or R_{ST}. However, the higher ploidy means that the overall size of the complete distance matrix increases. So what is needed to adapt a standard diploid AMOVA to autopolyploid data is to account for this larger overall sample size in all the calculations, which is very straightforward when the data contain only a single ploidy level. For a total sample size of N diploids, the distance matrix is of size $2N{*}2N$, whereas for autopolyploids with ploidy level x, the matrix is of size $xN{*}xN$ (for computational efficiency it is also possible to only use the lower or upper half of the matrix). The Sums of Squares are therefore calculated by summing over a larger number of pairwise distances, though this follows the same approach as outlined by Excoffier et al. (1992; Their equations 8a—8c) by summing over groups, populations, and individuals as necessary.

The higher ploidy level results in a larger number of allele copies within individuals, populations, and groups. This larger number of allele copies needs to be reflected in the degrees of freedom used for calculating the Expected Mean Squares; however, this is only the case for the within-individual and total degrees of freedom, as the others are only determined by the higher-level sample sizes. In **Table 1** we give generic formulas for the degrees of freedom for any ploidy $x > 1$ for a model with a single group of populations, and compare this to the diploid case described in the original papers (Excoffier et al., 1992; Peakall et al., 1995; Michalakis and Excoffier, 1996); the notation follows the notation used in the documentation of the software Arlequin (Excoffier and Lischer, 2010, 2015). **Table 2** shows the same, but then for multiple groups of populations. The Expected Mean Squares can then be obtained by dividing the corresponding Sum of Squares by the degrees of freedom.

For calculating the covariance components from the expected mean squares it is necessary to incorporate the sample sizes at the different hierarchical levels included in the analysis. The simplest case is a single group of populations all of the same ploidy level x and all with the same sample size N_p. In this case (**Table 1**), the multiplication factor n is defined as xN_p, the number of allele copies sampled per population. However, when there is unbalanced sampling (**Table 1**), this multiplication factor has to take the sample sizes for all populations separately into account:

$$n = \frac{xN - \sum_{p\in P} \frac{xN_p^2}{N}}{P-1} \tag{3}$$

Here, N_p is the number of individuals sampled in population p.

When there are multiple groups of populations (**Table 2**) there are three coefficients: n, n', and n''. When sampling is balanced, so with the same number of individuals sampled for every population and the same number of populations sampled in each of the G groups, n and n' are defined as xN_p. As above, this is simply the number of allele copies sampled per population. The value of n'' is then defined as xN_g, the number of allele copies sampled per group of populations. However when sample sizes within populations and/or groups are unbalanced (**Table 2**), the sample sizes have to be taken into account for the calculation, and the three coefficients are defined as:

$$n = \frac{xN - \sum_{g\in G}\sum_{p\in g} \frac{xN_p^2}{N_g}}{P-G} \tag{4a}$$

$$n' = \frac{\sum_{g\in G} \frac{(N-N_g)}{N_g} \sum_{p\in g} xN_p^2}{N(G-1)} \tag{4b}$$

$$n'' = \frac{xN - \frac{\sum_{g\in G} xN_g^2}{N}}{G-1} \tag{4c}$$

TABLE 1 | Outline of the AMOVA framework for a single group of populations with the degrees of freedom (d.f.) both given for diploids and generalized for any ploidy level *x* (except haploid).

Source	d.f. diploid	d.f. *x*-ploid	Sum of squares	Expected mean squares
Among populations	*P-1*	*P-1*	SSD(*AP*)	$n\sigma_a^2+x\sigma_b^2+\sigma_c^2$
Among individuals within populations	*N-P*	*N-P*	SSD(*AI/WP*)	$x\sigma_b^2+\sigma_c^2$
Within individuals	*N*	*(x-1)·N*	SSD(*WI*)	σ_c^2
Total	*2N-1*	*x·N-1*	SSD(*T*)	σ_T^2

P is the number of populations and N the number of individuals; the value of the multiplication coefficient n is calculated using Equation (3). This method can be used to obtain estimates of F_{ST}, φ_{ST}, or R_{ST}.

TABLE 2 | Outline of the AMOVA framework for multiple group of populations with the degrees of freedom (d.f.) both given for diploids and generalized for any ploidy level *x* (except haploid).

Source	d.f. diploid	d.f. *x*-ploid	Sum of squares	Expected mean squares
Among groups	*G-1*	*G-1*	SSD(*AG*)	$n''\sigma_a^2+n'\sigma_b^2+x\sigma_c^2+\sigma_d^2$
Among populations within groups	*P-G*	*P-G*	SSD(*AP/WP*)	$n\sigma_b^2+x\sigma_c^2+\sigma_d^2$
Among individuals within populations	*N-P*	*N-P*	SSD(*AI/WP*)	$x\sigma_c^2+\sigma_d^2$
Within individuals	*N*	*(x-1)·N*	SSD(*WI*)	σ_d^2
Total	*2N-1*	*x·N-1*	SSD(*T*)	σ_T^2

G is the number of groups, P the number of populations, and N the number of individuals; the value of the multiplication coefficients n, n′ and n″ are calculated using Equations (4a–c). This method can be used to obtain estimates of F_{ST}, φ_{ST}, or R_{ST}.

where N_g is the number of individuals sampled in group g. For haploid and diploid data ($x = 1$ and $x = 2$), these equations are the same as for the standard AMOVA (Michalakis and Excoffier, 1996; Excoffier and Lischer, 2015).

Mixed Ploidy Datasets

Slightly more complicated evolutionary scenarios involve multiple ploidy levels, either occurring in separate populations, or co-occurring in populations. In such a case, there is no single ploidy level x that can be used to calculate the degrees of freedom and the multiplication coefficients. However, when the ploidy level of every genotyped individual is known (e.g., through flow cytometry), this problem can be solved by using the number of allele copies sampled per population (C), rather than the number of individuals (N). **Table 3** shows the formulas for the degrees of freedom for any mixture of ploidy levels (though all should be at least diploids) for a model with a single group of populations. The corresponding coefficients n, n', and n'' are defined as (again

TABLE 3 | Outline of the AMOVA framework for a single group of populations with the degrees of freedom (d.f.) for a mixture of individuals with different ploidy levels, based on the total number of allele copies sampled (*C*).

Source	d.f. *x*-ploid	Sum of squares	Expected mean squares
Among populations	*P-1*	SSD(*AP*)	$n''\sigma_a^2+n'\sigma_b^2+\sigma_c^2$
Among individuals within populations	*N-P*	SSD(*AI/WP*)	$n\sigma_b^2+\sigma_c^2$
Within individuals	*C-N*	SSD(*WI*)	σ_c^2
Total	*C-1*	SSD(*T*)	σ_T^2

P is the number of populations and N the number of individuals; the value of the multiplication coefficients n, n′, and n″ are calculated using Equations (5a–d). This method can be used to obtain estimates of F_{ST}, φ_{ST}, or R_{ST}.

following the notation from Excoffier and Lischer, 2015):

$$S_P = \sum\nolimits_{p\in P}\sum\nolimits_{i\in p}\frac{C_i^2}{C_p} \tag{5a}$$

$$n = \frac{C-S_P}{N-P} \tag{5b}$$

$$n' = \frac{S_P-\sum_{i\in N}\frac{C_i^2}{C}}{P-1} \tag{5c}$$

$$n'' = \frac{C-\sum_{p\in P}\frac{C_p^2}{C}}{P-1} \tag{5d}$$

The *F*-statistics can then be calculated in the normal way, using Equations (1a–c). Note that when the significance of the population differentiation is tested by permuting individuals over populations, the number of allele copies in the permuted populations may differ from the original values. Therefore, the coefficients n, n', and n'' will have to be recalculated for every permutation.

Ploidy-Independent ρ-Statistic

In addition to the above-developed method that yields estimates of F_{ST}, φ_{ST}, or R_{ST}, AMOVA can also be used to obtain estimates of the ploidy-independent ρ-statistic (Ronfort et al., 1998). Here we show that this can be done by performing AMOVA on a matrix of squared Euclidean distances calculated from the within-individual allele frequencies. Other than the above methods of calculating distances—where each distance is calculated between a pair of alleles or haplotypes—here each squared Euclidean distance (denoted as d_{ij}^2) is calculated between a pair of individual genotypes at a locus. The metric is calculated as

$$d_{ij}^2 = \sum\nolimits_{a=1}^{A}\left(p_{ia}-p_{ja}\right)^2 \tag{6}$$

where p_{ia} is the frequency of the ath allele ($a \in \{1, 2, \ldots, A\}$) within individual i. In diploids, these frequencies can take the values 0, 0.5, and 1; in triploids the values 0, 0.33, 0.67, and 1; in tetraploids the values 0, 0.25, 0.5, 0.75, and 1; etc. For haploids, the only two possible values are 0 and 1 and therefore for haploids this metric is the same as the simple-matching distance; by extension this means that for haploid data the value of ρ equals that of F_{ST}.

This distance metric yields, for any ploidy level, only a single distance value per pair of individuals. As a result, the distance matrix is only of size N^*N, whereas the approach above resulted in a matrix of xN^*xN, for data of ploidy level x. The N^*N matrix can then be used to perform AMOVA using the equations (not shown here) originally developed for haploid data in the paper by Excoffier et al. (1992). This approach also allows ρ to be calculated at different hierarchical levels, e.g., to compare differentiation among clusters of populations. For such use, we will adopt the convention of adding subscripts to indicate which levels are compared, though Ronfort et al. (1998) did not use any such subscripts in their original description of ρ. Note that since the within-individual component is disregarded, there are no ρ equivalents of F_{IS} and F_{IT} in such a hierarchical analysis.

When the two individuals have the same ploidy level, the squared Euclidean distance metric proposed here is a simple linear transformation of the squared Euclidean distance metric of Smouse and Peakall (1999). Since a linear transformation of the distance matrix does not affect the relative sizes of the variance components, this means that the Smouse and Peakall distance can also be used for AMOVA. However, the metric from Smouse and Peakall has only been defined for cases where the two individuals have the same ploidy level, whereas the metric proposed above is also suited to mixtures of different ploidy levels.

The mathematical relationship between the squared Euclidean distance metric and ρ can be deduced as follows. Again, p_{ia} refers to the frequency of the ath allele ($a \in \{1, 2, \ldots, A\}$) in the ith individual ($i \in \{1, 2, \ldots, N\}$). The sum of the **D** matrix can then be transformed as:

$$\begin{aligned}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}d_{ij}^2 &= \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\sum\nolimits_{a=1}^{A}\left(p_{ia}-p_{ja}\right)^2\\ &= 2N^2\sum\nolimits_{a=1}^{A}\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}p_{ia}^2\right.\\ &\qquad\left.-\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}p_{ia}\right)^2\right)\end{aligned} \tag{7}$$

If we define

$$\check{H}_O \equiv \frac{1}{N}\sum\nolimits_{i=1}^{N}\left(1-\sum\nolimits_{a=1}^{A}p_{ia}^2\right) \tag{8}$$

and

$$\check{H}_E \equiv 1-\sum\nolimits_{a=1}^{A}\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}p_{ia}\right)^2 \tag{9}$$

then the sum of squared distances in Equation (7) can be simplified to:

$$\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}d_{ij}^2 = 2N^2\left(\check{H}_E-\check{H}_O\right) \tag{10}$$

$\check{H}_E$ and $\check{H}_O$ as defined here are analogous—but not equivalent—to the standard H_E and H_O as defined by Nei

(1987) for diploids and by Moody et al. (1993) for polyploids (see also Meirmans et al., 2018):

$$H_O = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\left(1-\sum\nolimits_{a=1}^{A}p_{ia}^2\right)\cdot\frac{x_i}{x_i-1}\right) \tag{11}$$

$$H_E = 1-\sum\nolimits_{a=1}^{A}\left(\frac{\sum_{i=1}^{N}\left(x_i\cdot p_{ia}\right)}{\sum_{i}^{N}x_i}\right)^2 \tag{12}$$

While H_E and H_O attempt to correct the calculation of allele frequency or heterozygosity using individual ploidy information, $\check{H}_E$ and $\check{H}_O$ ignore such information, hence endowing ρ a ploidy-independent nature.

In the Island model where the number of populations is r (each population has a size of N), ρ_{ST} can be calculated as

$$\rho_{ST} = 1-\frac{(r\cdot N)^2\cdot x^2\sum_{k=1}^{r}\sum_{i=1}^{N}\sum_{j=1}^{N}d_{kij}^2}{r\cdot N^2\cdot x^2\sum_{i=1}^{r\cdot N}\sum_{j=1}^{r\cdot N}d_{ij}^2} \tag{13}$$

Using the link between the sum of squared distances and the $\check{H}$-statistics that was established in Equation (10), Equation (13) can be transformed into:

$$\rho_{ST} = \frac{\check{H}_T-\check{H}_S}{\check{H}_T-\check{H}_O} \tag{14}$$

The statistic $\check{H}_T$ is defined in the same vein as $\check{H}_E$ but then for all populations together; $\check{H}_S$ is the average of $\check{H}_E$ calculated over populations. If the populations contain only a single ploidy level, Equation (14) can be transformed into

$$\rho_{ST} = \frac{H_T-H_S}{H_T-H_O\cdot\frac{x-1}{x}} \tag{15}$$

which is the same as Equation (6) in Meirmans et al. (2018).

APPLICATION TO DATA

Simulations Under a Hierarchical Island Model

To test how well the above-developed AMOVA framework performs for data with different ploidy levels, we simulated data under a standard hierarchical island model (Slatkin and Voelm, 1991; Vigouroux and Couvet, 2000). A set of 20 populations was simulated, divided into two archipelagoes, both having 10 populations. All populations had the same size of $N = 100$; mating within populations was completely random, including a probability of self-fertilization of $1/N$. Genetic markers were simulated at 1,000 independently segregating loci; mutation followed a K-alleles model with 100 possible allelic states and a mutation rate of $\mu = 0.0001$. Migration took place at different rates among populations from the same archipelago (*m1*) and among populations from different archipelagoes (*m2*).

The model was population-based, so individuals were not explicitly modeled but instead the populations were represented by a set of vectors containing the allele frequencies of all possible allelic states at all loci. Under the assumption of random mating, one generation of genetic drift can then easily be simulated by drawing random numbers from a multinomial distribution. For the expected values in the multinomial, we used the current population allele frequencies—after incorporating the expected effects of migration and mutation. For the number of draws in the multinomial, we used the number of chromosome copies in the population, so the population size multiplied by the ploidy level. The model was written in *R*, using the *rmultinom*() function for drawing random numbers; the used *R*-script is available in online Supplement 1 (Data Sheet 1).

The model was run for diploids, tetraploids, and hexaploids, for values of *m2* of 0.001, 0.0001, and 0.00001; per value of *m2* a range of values of *m1* was used with a maximum of 0.1 and a minimum equal to the value of *m2* (so *m1* ≥ *m2*). Per scenario, the model was run once for 20,000 generations; replication was provided by the use of the 1,000 independent loci. After the last generation, genotypes were constructed by randomly distributing the alleles over individuals and written to a file. The software GENODIVE v. 2b27 (Meirmans and van Tienderen, 2004) was used to perform a hierarchical AMOVA on the resulting genotypes. The results were compared to the theoretical expectations for F_{SC} and F_{CT} derived by Vigouroux and Couvet (2000). Though these expectations were only derived for diploids, general results for any ploidy level x can be obtained by substituting all occurrences of the term "*4N*" in the equations by the term "*2xN*" (see Meirmans et al., 2018). The expectations for ρ for any ploidy level are equivalent to the expectation for F_{ST} under haploidy (Ronfort et al., 1998; Meirmans et al., 2018), so can also be derived from the equations of Vigouroux and Couvet (2000) by substituting every "*4N*" by "*2N*."

Simulation Results

When applying the AMOVA framework to the simulated data for several ploidy levels, the results closely matched the theoretical expectations (**Figure 1**), indicating that AMOVA correctly estimates the variance components and the F-statistics. For all three values of *m2*, F_{SC} showed a monotonic decrease with increasing values of *m1* (**Figure 1**, top row), whereas F_{CT} showed a monotonic increase (**Figure 1**, bottom row). As random mating within populations was assumed, the values of F_{IS} were close to zero for all simulated scenarios (not shown). The only slight deviation between the results of the simulation and the theoretical expectations was observed for the F_{CT}-statistic when the migration rate within archipelagoes (*m1*) was close to or equal to the migration rate between archipelagoes (*m2*). This deviation can easily be explained since the theoretical derivations of Vigouroux and Couvet (2000) assume that *m1* > *m2*. For the cases where *m1* = *m2*, the simulations consistently show a F_{CT} value that is close to zero, whereas the expected values are slightly higher.

As expected, there is a strong difference between the F-statistics (**Figure 1**) and the ρ-statistics (**Figure 2**) in how they behave under different ploidy levels. For the F-statistics, at a given set of migration rates, the values decrease with increasing ploidy level. This is due to the increased impact of migration at higher ploidy levels combined with a decrease in the force of genetic drift (Meirmans et al., 2018). On the other hand, the ρ-statistics generally have similar values for all ploidy levels when

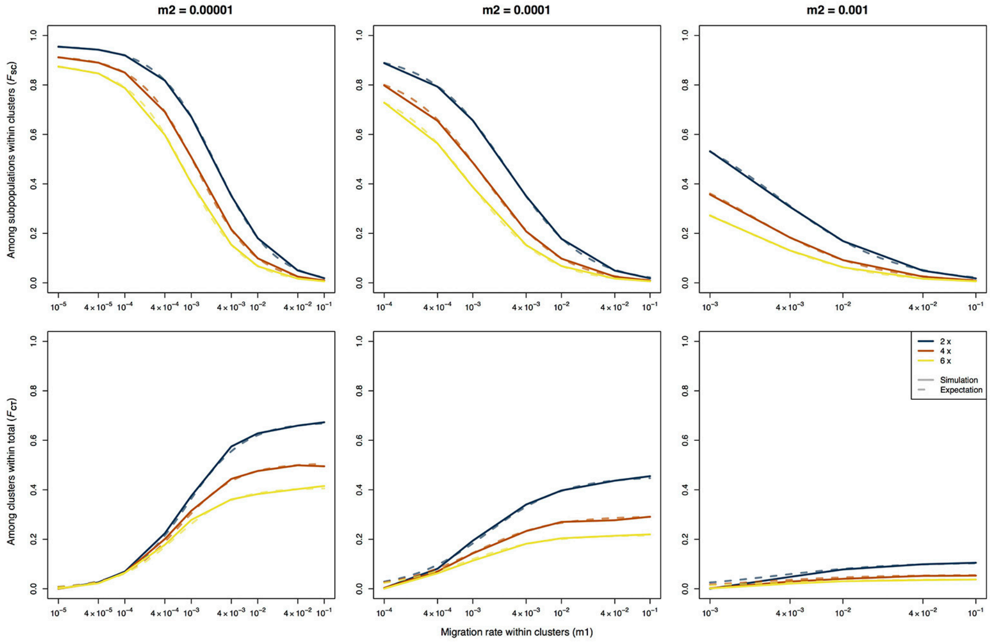


FIGURE 1 | *F*-statistics calculated using an AMOVA on data of three different ploidy levels simulated using a hierarchical island model of migration. The solid lines represent the results of the simulations, the dashed lines represent the expected values based on the derivations of Vigouroux and Couvet (2000).

calculating the differentiation among subpopulations within clusters (ρ_{SC}) or the differentiation among clusters (ρ_{CT}). This ploidy-independence of the ρ-statistic is immediately evident from the almost completely overlapping lines in **Figure 2**. As we saw above for F_{CT}, the estimates of ρ_{CT} from the simulated data show a slight deviation from the expected values when the assumption of $m1 > m2$ is violated.

DISCUSSION

Expanding the AMOVA Framework

In this paper, we showed how the AMOVA framework (Excoffier et al., 1992; Peakall et al., 1995; Michalakis and Excoffier, 1996) can be used for autopolyploids of any ploidy level by adapting the way the Sums of Squares and resulting variance components are calculated. This method can be used with any distance metric that is normally used with haploid or diploid data, which means that the method can be used to obtain estimates of F_{ST}, φ_{ST}, or R_{ST}. In addition, we showed that the use of a simple squared Euclidean distance metric defined here will yield an estimate of the ploidy-independent ρ-statistic. For both approaches (F_{ST} and ρ), AMOVA can be used for datasets from a single cytotype or a mixture of cytotypes. Since the covariance components are calculated separately for each locus, the method can even be used with species where there is ploidy variation within the genome, such as the salmonid fishes (Allendorf et al., 2015).

We tested the developed method with datasets simulated under a hierarchical island model of migration, for multiple ploidy levels. The results of the simulations closely matched those from the theoretical derivation of Vigouroux and Couvet (2000; see also Slatkin and Voelm, 1991), showing that the method correctly estimates the variance components. A slight deviation was only observed when the assumption of $m1 > m2$ that was made by Vigouroux and Couvet was violated. The violation of this assumption was done on purpose as the simulations where $m1 = m2$ allowed us to test the AMOVA in scenarios without any hierarchical population structure. In these cases, the AMOVA correctly showed the absence of any differentiation between clusters ($F_{CT} = 0$); the theoretical expectation in these cases was slightly higher. Interestingly, this is the first study—as far as we are aware—that has compared the theoretical expectations for the hierarchical island model with simulated data; even though hierarchical *F*-statistics are widely used in analyses of genetic marker data, the theoretical derivations have received very little attention, for autopolyploids as well as for diploids.

The Ploidy-Independent ρ-Statistic

Though the ρ-statistic that was developed by Ronfort et al. (1998) is ideally suited to analyze autopolyploid data, it has seen relatively little use for this purpose. We hope that the possibility of calculating ρ using AMOVA will help to make it more widely adapted. For calculating ρ we described a simple squared Euclidean distance metric based on within-individual allele frequencies. This is closely related to the metric of Smouse and Peakall (1999), which uses allele counts rather than frequencies. As we describe above, for any single ploidy level our metric is a simple linear transformation of the metric of Smouse and Peakall, and so for a single-ploidy dataset the two metrics give identical results in AMOVA. However, one problem with the Smouse and Peakall metric—and AMOVA based on it—is that it

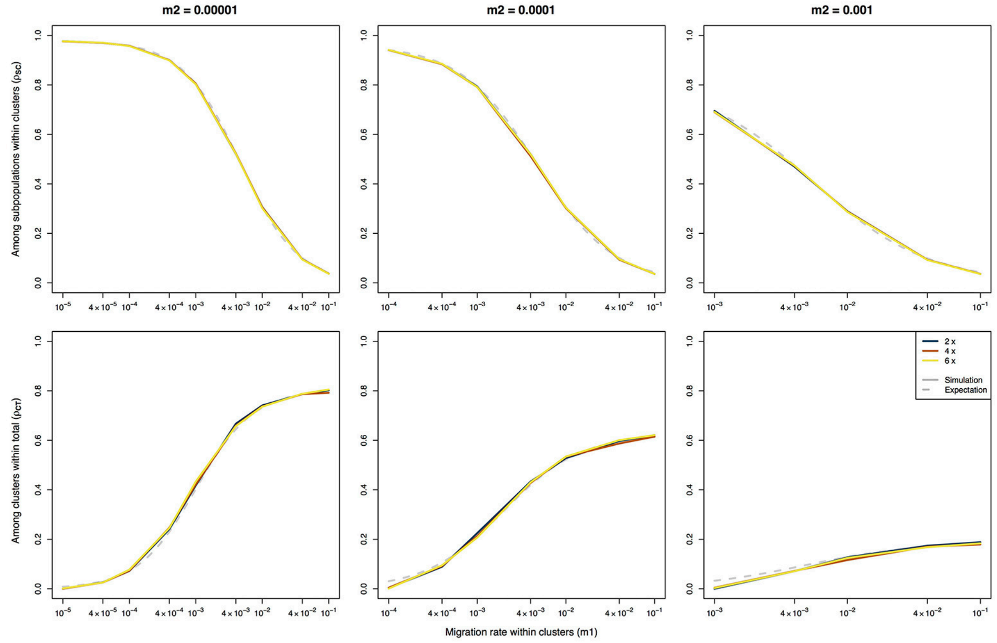


FIGURE 2 | Ploidy-independent ρ-statistics calculated using an AMOVA on data of three different ploidy levels simulated using a hierarchical island model of migration. The solid lines represent the results of the simulations, the dashed lines represent the expected values based on the derivations of Vigouroux and Couvet (2000). Note that the expectations are the same for all three ploidy levels.

cannot be used for analyses with mixed ploidy levels, as that will lead to a bias. The metric from Smouse and Peakall (1999) has received some criticism because it is founded on geometric rather than biological principles (Kosman and Leonard, 2005; Dufresne et al., 2014). However, these criticisms are unjustified since our deductions above (Equations 7–15) recovered the biological meaning of this method by linking the metric with the calculation of ρ.

In our derivations above we have only focused on autopolyploids. However, in many polyploid species it is not known whether it is an allopolyploid or an autopolyploid. In addition, species may show inheritance patterns that are intermediate between these two extremes, with partly disomic and partly polysomic inheritance (segmental allopolyploids; Stebbins, 1947). Furthermore, the frequency of polysomic inheritance may even vary among loci within a genome (Stift et al., 2008). Meirmans and van Tienderen (2013) used simulations of tetraploids where the rate of tetrasomy varied between full disomic and full tetrasomic inheritance to test the presence of bias in several genetic summary statistics. They found that an assumption of autopolyploidy for a species that is in fact an allopolyploid can give a strong downward bias in the value of F_{ST}. On the other hand, the ρ-statistic was almost completely free of such a bias and is therefore the statistic of choice when the exact mode of segregation of a polyploid is unknown. Of course, this does not mean that the mode of segregation becomes irrelevant for the analysis of polyploid data; for a true understanding of the genetic processes within a polyploid species, studying the segregation mode is indispensible.

The greatest strength of ρ lies in comparisons across species or sets of populations with different ploidy levels. For a given migration rate, population size, and mutation rate, the value of ρ will be the same in diploids as in polyploids. Comparisons of ρ across species with different ploidy levels therefore permits assessing whether the impact of these processes are different in the different species. The ρ-statistic can also be easily calculated for mixed-ploidy data. However, in such cases there is an important caveat. Whereas the same set of allele frequencies always yields the same value of F_{ST}, regardless of the ploidy level, this is not the case for ρ. So in a case where there are multiple ploidy levels, calculating ρ separately for each ploidy level will give different values, even when within populations there is complete admixture among the cytotypes (see Meirmans et al., 2018). Another limitation of ρ is that is currently only defined under the Infinite Allele and K-allele models of mutation. This means that it is not applicable to markers that follow a Stepwise Mutation Model (as is the case for R_{ST}) or for sequence data (as is the case for φ_{ST}). This is because there are no Euclidean distances among individual genotypes that can take these mutational processes into account.

Polyploid AMOVA in Practice: Software

AMOVA for autopolyploids has been implemented in the software GenoDive (Meirmans and van Tienderen, 2004), which is freely available for Mac computers from http://www.patrickmeirmans.com/software. In addition to F_{ST} and ρ, GenoDive can also use AMOVA to calculate the F'_{ST} statistic for autopolyploids, which is F_{ST} standardized relative to the level of within-population variation (Meirmans, 2006;

Meirmans and Hedrick, 2011). Besides the standard AMOVA, where the degree of differentiation is calculated based on an *a priori* defined hierarchical population structure, GENODIVE also offers AMOVA-based *K*-means clustering (Meirmans, 2012) for autopolyploid data based on the ρ-statistic. This analysis allows clustering of individuals or populations into *k* groups, where the algorithm finds the clustering with the highest value of the ρ-statistic. The autopolyploid AMOVA has also been implemented in the R-package POPPR V. 2.7.0 (Kamvar et al., 2014).

The ρ-statistic is also applicable to haploid and diploid data, and for such datasets it can be estimated using AMOVA with the software GENALEX (Peakall and Smouse, 2012). For haploid data, the ρ-statistic is simply equal to F_{ST} obtained from running an AMOVA. For diploid data, the option to calculate genetic distances among individuals should first be run, which calculates the metric of Smouse and Peakall (1999). When an AMOVA is subsequently performed using this distance matrix, the resulting differentiation statistics—labeled φ in the output—are equivalent to ρ.

Dealing With Missing Dosage Information

One of the major practical challenges of working with autopolyploids is the problem of missing dosage information for alleles (Dufresne et al., 2014). Depending on the type of marker—and the sequencing depth for genotyping-by-sequencing data—often only marker phenotypes are available and not the complete genotypes. This missing dosage information may cause a bias in the estimation of allele frequencies in samples from autopolyploid populations; in AMOVA, this will cause a bias in the estimation of the covariance components. This is because individuals with different genotypes can have the same phenotype: *AAAB*, *AABB*, and *ABBB* all have phenotype *AB*. This will lead to an underestimation of the distance between individuals and the corresponding Sums of Squares, and hence to an underestimation of F_{ST} and ρ. It is, as yet, not possible to correct for this bias directly in the calculation of AMOVA.

It is possible to correct for this bias in an indirect way by completing the genotypes via random imputation of the missing alleles, when Hardy-Weinberg equilibrium can be assumed within populations. For this, bias-corrected allele frequencies should first be estimated based on the set of phenotypes, e.g., using the maximum likelihood method of De Silva et al. (2005). Then for every individual, the phenotype should be filled in by randomly drawing alleles based on the expected frequency (under HWE) of the different genotypes that can be constructed from this phenotype, given the estimated frequencies of the alleles present in the phenotype. So for example when a tetraploid has phenotype *AB* and allele *A* is very common in the population and allele *B* is very rare, it is much more likely that the genotype will be randomly filled to *AAAB* than to *AABB* or *ABBB*. If this imputation is done for all individuals in the dataset and the sample sizes per population are sufficient, the allele frequencies in the imputed dataset will closely match the estimated allele frequencies and the imputed dataset can be used to perform AMOVA. Simulations have shown that this type of imputation can successfully remove bias caused by missing dosage for both F_{ST} and ρ (Meirmans et al., 2018). The procedure has been implemented in the AMOVA and AMOVA-based *K*-means clustering functions of the software GENODIVE (Meirmans and van Tienderen, 2004). Since it involves randomly drawing alleles, it may be prudent to repeat the procedure a number of times and calculate the average values of the *F*-statistics across replicates. Nevertheless, it's important to realize that the assumption of random mating, necessary for such imputation, is likely to be violated for many polyploids. Therefore, a next major step in the field would be the development of a method that can take the missing dosage into account directly without an assumption of HWE.

CONCLUSIONS

The statistical tools available for polyploids still lag behind those available for diploids (Dufresne et al., 2014; Meirmans et al., 2018). Hopefully, the Analysis of Molecular Variance for autopolyploids that we described here will help to narrow this gap when developers of statistical software that allows polyploid data (e.g., Jombart, 2008; Clark and Jasieniuk, 2011; Kamvar et al., 2014) will implement this method more widely. We also hope that our description of the link between the squared Euclidean distances, calculated from the within-individual allele-frequencies, and the ρ-statistic will help advocate the use of this statistic. Its independence of the ploidy level, the rate of double reduction, the frequency of polysomic inheritance, and the mating system makes ρ better suited for comparisons among species than the standard F_{ST}, both for diploids and for polyploids.

AUTHOR CONTRIBUTIONS

PM: developed the general method for calculating the Sums of Squares for polyploids; SL: derived the proof linking the ρ-statistic to the Euclidean distances; PM: wrote the manuscript with input from SL.

ACKNOWLEDGMENTS

We would like to thank Marc Stift, Filip Kolář, and the participants of the 2017 polyploidy workshop in Prague for stimulating discussions about all aspects of polyploidy research. We also would like to thank Zhian Kamvar for implementing the polyploid AMOVA in the POPPR package.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2018.00066/full#supplementary-material

REFERENCES

Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W., and Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J. Hered.* 106, 217–227. doi: 10.1093/jhered/esv015

Bever, J. D., and Felber, F. (1992). The theoretical population genetics of autopolyploidy. *Oxford Surv. Evol. Biol.* 8, 185–217.

Clark, L. V., and Jasieniuk, M. (2011). POLYSAT: an R package for polyploid microsatellite analysis. *Mol. Ecol. Resour.* 11, 562–566. doi: 10.1111/j.1755-0998.2011.02985.x

Cockerham, C. (1973). Analysis of gene frequencies. *Genetics* 74, 679–700.

De Silva, H. N., Hall, A. J., Rikkerink, E., McNeilage, M. A., and Fraser, L. G. (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity* 95, 327–334. doi: 10.1038/sj.hdy.6800728

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Dupanloup, I., Schneider, S., and Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.* 11, 2571–2581. doi: 10.1046/j.1365-294X.2002.01650.x

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Excoffier, L., and Lischer, H. E. L. (2015). *Arlequin Ver 3.5: An Integrated Software Package for Population Genetics Data Analysis.* Available online at: http://cmpg.unibe.ch/software/arlequin35

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial-DNA restriction data. *Genetics* 131, 479–491.

Excoffier, L., and Smouse, P. E. (1994). Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics* 136, 343–359.

Field, D. L., Broadhurst, L. M., Elliott, C. P., and Young, A. G. (2017). Population assignment in autopolyploids. *Heredity* 119, 389–401. doi: 10.1038/hdy.2017.51

Glennon, K. L., Ritchie, M. E., and Segraves, K. A. (2014). Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecol. Lett.* 17, 574–582. doi: 10.1111/ele.12259

Hardy, O. J. (2015). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Mol. Ecol. Resour.* 16, 103–117. doi: 10.1111/1755-0998.12431

Hardy, O. J., and Vekemans, X. (2002). SPAGEDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Jombart, T. (2008). ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). POPPR: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281. doi: 10.7717/peerj.281

Kolár, F., Certner, M., Suda, J., Schönswetter, P., and Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Plant Sci.* 22, 1041–1055. doi: 10.1016/j.tplants.2017.09.011

Kosman, E., and Leonard, K. J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Mol. Ecol.* 14, 415–424. doi: 10.1111/j.1365-294X.2005.02416.x

Larkin, K., Tucci, C., and Neiman, M. (2016). Effects of polyploidy and reproductive mode on life history trait expression. *Ecol. Evol.* 6, 765–778. doi: 10.1002/ece3.1934

Li, C. (1976). *Population Genetics*. Pacific Grove, CA: Boxwoord Press.

Meirmans, P. G. (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* 60, 2399–2402. doi: 10.1111/j.0014-3820.2006.tb01874.x

Meirmans, P. G. (2012). AMOVA-based clustering of population genetic data. *J. Hered.* 103, 744–750. doi: 10.1093/jhered/ess047

Meirmans, P. G., and Hedrick, P. (2011). Assessing population structure: F_{ST} and related measures. *Mol. Ecol. Resour.* 11, 5–18. doi: 10.1111/j.1755-0998.2010.02927.x

Meirmans, P. G., and van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* 4, 792–794. doi: 10.1111/j.1471-8286.2004.00770.x

Meirmans, P. G., and van Tienderen, P. H. (2013). The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* 110, 131–137. doi: 10.1038/hdy.2012.80

Meirmans, P. G., Liu, S., and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Hered.* 109, 283–296. doi: 10.1093/jhered/esy006

Michalakis, Y., and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142, 1061–1064.

Moody, M. E., Mueller, L. D., and Soltis, D. E. (1993). Genetic variation and random drift in autotetraploid populations. *Genetics* 134, 649–657.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York, NY: Columbia University Press.

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

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

Peakall, R., Smouse, P. E., and Huff, D. R. (1995). Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss (*Buchloë dactyloides* (Nutt.) Engelm.). *Mol. Ecol.* 4, 135–147 doi: 10.1111/j.1365-294X.1995.tb00203.x

Ronfort, J., Jenczewski, E., Bataillon, T., and Rousset, F. (1998). Analysis of population structure in autotetraploid species. *Genetics* 150, 921–930.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139, 457–462.

Slatkin, M., and Voelm, L. (1991). F_{ST} in a hierarchical island model. *Genetics* 127, 627–629.

Smouse, P. E., and Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* 82, 561–573. doi: 10.1038/sj.hdy.6885180

Soltis, D. E., Soltis, P. S., Schemske, D. W., Hancock, J. F., Thompson, J. N., Husband, B. C., et al. (2007). Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56, 13–30. doi: 10.2307/25065732

Stebbins, G. L. (1947). Types of polyploids; their classification and significance. *Adv. Genet.* 1, 403–429. doi: 10.1016/S0065-2660(08)60490-3

Stift, M., Berenos, C., Kuperus, P., and van Tienderen, P. H. (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics* 179, 2113–2123. doi: 10.1534/genetics.107.085027

Vigouroux, Y., and Couvet, D. (2000). The hierarchical island model revisited. *Genet. Sel. Evol.* 32, 395–402. doi: 10.1186/1297-9686-32-4-395

Weir, B. S., and Cockerham, C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.

Whitlock, M. C. (2011). G'_{ST} and D do not replace F_{ST}. *Mol. Ecol.* 20, 1083–1091. doi: 10.1111/j.1365-294X.2010.04996.x

Wright, S. (1969). *Evolution and the Genetics of Populations, Vol. 2, The Theory of Gene Frequencies*. Chicago, IL: University Chicago Press.

Conflict of Interest Statement: 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: 23 October 2018
doi: 10.3389/fgene.2018.00480

The Site Frequency/Dosage Spectrum of Autopolyploid Populations

Luca Ferretti[1]*, *Paolo Ribeca*[1] and *Sebastian E. Ramos-Onsins*[2]

[1] The Pirbright Institute, Woking, United Kingdom, [2] Centre for Research in Agricultural Genomics, Barcelona, Spain

The Site Frequency Spectrum (SFS) and the heterozygosity of allelic variants are among the most important summary statistics for population genetic analysis of diploid organisms. We discuss the generalization of these statistics to populations of autopolyploid organisms in terms of the joint Site Frequency/Dosage Spectrum and its expected value for autopolyploid populations that follow the standard neutral model. Based on these results, we present estimators of nucleotide variability from High-Throughput Sequencing (HTS) data of autopolyploids and discuss potential issues related to sequencing errors and variant calling. We use these estimators to generalize Tajima's *D* and other SFS-based neutrality tests to HTS data from autopolyploid organisms. Finally, we discuss how these approaches fail when the number of individuals is small. In fact, in autopolyploids there are many possible deviations from the Hardy–Weinberg equilibrium, each reflected in a different shape of the individual dosage distribution. The SFS from small samples is often dominated by the shape of these deviations of the dosage distribution from its Hardy–Weinberg expectations.

Keywords: autopolyploidy, dosage distribution, Hardy-Weinberg equilibrium, high-throughput sequencing, site frequency spectrum, heterozygosity, neutrality tests, allelic dosage

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Barbara K. Mable,
University of Glasgow,
United Kingdom
Paul David Blischak,
The Ohio State University,
United States
Polina Yu. Novikova,
VIB-UGent Center for Plant Systems
Biology, Belgium

***Correspondence:**
Luca Ferretti
luca.ferretti@gmail.com

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

Received: *20 April 2018*
Accepted: *28 September 2018*
Published: *23 October 2018*

Citation:
Ferretti L, Ribeca P and
Ramos-Onsins SE (2018) The Site
Frequency/Dosage Spectrum of
Autopolyploid Populations.
Front. Genet. 9:480.
doi: 10.3389/fgene.2018.00480

1. INTRODUCTION

The study of nucleotide variability in polyploid species is a convoluted task that requires solving a number of methodological and analytical difficulties related to the specific nature of the species (detailed in the reviews of Dufresne et al., 2014; Meirmans et al., 2018). The impact of diploidy on the evolutionary dynamics is well-known, but the complexity of the impact of higher ploidy on the genetic variability of polyploid organisms is even higher. An example is provided by autopolyploid species: as they contain copies originating from genome duplication of the same species, the inheritance is expected to be polysomic (all the variants of the same chromosome can pair in the meiosis process) but it is not rare to find preferential pairs (Stift et al., 2008; Chester et al., 2012), resulting in partial polysomic or even disomic inheritance. The different inheritance types, which may simultaneously occur in the same species, could generate differences in the effective population size at different loci and consequently different patterns of genetic variability. Another distinctive aspect of polyploid species that impacts their genetic variability patterns is the process of *double reduction*, where the two copies of the same chromatid migrate to the same gamete (Haldane, 1930). As a consequence, this process will increase drastically the homozygosity of the gametes for the involved segment.

High-Throughput Sequencing (HTS) has facilitated the study of genome data in general and that of polyploid species as well. Still there are difficulties, mainly assigning the sequence reads to homologous (rather than homeologous) loci and/or dealing with relatively high rates of sequencing error (You et al., 2018). The amount of software available in order to correctly assembly and detect variants (e.g., GATK from Broad Institute) is increasing, although the task remains challenging (Mielczarek and Szyda, 2016; You et al., 2018). These methodological problems are expected to be (at least partially) solved in the next years with the technological progress of the sector, including long reads and linked reads to improve phasing and increased throughput of sequencing runs (Dufresne et al., 2014; Shendure et al., 2017).

The study of polyploid variability from HTS data and the development of statistical methods based on these sequencing methodologies are driving current genetic studies of polyploids (Dufresne et al., 2014; Hardy, 2016) and will continue to have a fundamental impact on the field. Nevertheless, still much work is needed, especially on the topic of allelic dosage, that is, the number of copies of each allele in a heterozygous individual (Blischak et al., 2016). Since the development of HTS, a number of studies developing computational and statistical methods that account for polyploidy have been published. Example are statistics to estimate the levels of variability (Ferretti and Ramos-Onsins, 2015) and heterozygosity (e.g., De Silva et al., 2005; Hardy, 2016) with different approaches to take into account the allelic dosage, or the detection of population structure (e.g., Falush et al., 2003; Gao et al., 2007) and comparative measures of these differences between populations/species/individuals (e.g., Jost, 2008; Meirmans and Hedrick, 2011). Arnold et al. (2012) showed that autotetrasomic inheritance can be modeled using a Kingman's standard coalescent (Kingman, 1982). Their results can be generalized to autopolyploid species of different ploidy and are especially useful as a null model to predict the neutral patterns of genetic diversity in polyploid species. Also additional phenomena specific to polyploids, such as *double reduction*, can be modeled in a way resembling partial self-fertilization (Arnold et al., 2012).

Nevertheless, a major gap in the population genetic analysis of polyploid organisms is the application of methods based on the Site Frequency Spectrum (SFS). Of special interest is the generalization to polyploid organisms of Tajima's *D* (Tajima, 1989), Fay and Wu's *H* (Fay and Wu, 2000) and other neutrality tests based on the SFS (Achaz, 2009; Ferretti et al., 2010, 2012). The SFS and the heterozygosity of allelic variants are among the most important statistics for population genetic analysis of diploid organisms and have been commonly used for describing the genetic variability of genomic data and for inferring the parameters of evolutionary models (e.g., Nielsen, 2000). Indeed, the combination of these two statistics (frequency and heterozygosity) describes completely the genotype of a diploid population for a given genomic position.

In this paper we consider a single population of autopolyploid organisms. Compared to the diploid case, the genotypes of variants in polyploid organisms present a more complex structure resulting from a combination of internal spectra for each individual. We discuss this genotype structure and its decomposition into different statistics, including the SFS and a generalization of the distribution of heterozygosity that we call the Site Dosage Spectrum (SDS).

For samples of large size, we argue that the details of deviations from Hardy–Weinberg equilibrium have a relatively small impact on the SFS. The expected value of the SFS of autopolyploid individuals is derived for a panmictic, neutral population of constant size. We also derive the expected value the most general spectrum for autopolyploids, i.e., the joint Site Frequency-Dosage Spectrum (SFDS), which represents a combination of the SFS and the SDS. We use these results as a null model to build estimators of nucleotide diversity and neutrality tests for HTS data and we discuss the robustness of estimators of genetic variability.

For small samples, violations of Hardy–Weinberg in the dosage distribution have a strong impact on the SFS. We show how autopolyploid populations have the potential to harbor a wide range of deviations from Hardy–Weinberg equilibrium due e.g., to inbreeding, population structure, selection, dominance, modes of inheritance, or combinations of these causes. We discuss the impact of some of these violations on dosage and on SFS-based neutrality tests.

A synopsis of symbols and abbreviations used in both text and formulas can be found in **Table 1**. It should be noted that to the best of our knowledge most of the equations that follow (all but 2, 3, 7, 11, and 13) are original work presented in this paper for the first time. More details about their derivations can be found in the **Appendix**.

2. SFDS STRUCTURE IN AUTOPOLYPLOIDS

2.1. SFS and Heterozygosity in Diploids

Individuals are often sampled from a wild population without prior studies of the subpopulation structure or phenotypic differences. In this case, it is usually assumed for population genetic analysis that all individuals are equivalent and that any summary statistic should treat all sequences equally. To

TABLE 1 | List of the main symbols and abbreviations used throughout the text.

Symbol	Meaning
p	Ploidy
n	Sample size
θ	Genetic variability, i.e., population-scaled mutation rate
ξ_j	Site Frequency Spectrum (SFS) for frequency j/n
d	Allelic dosage
$\mathcal{I}_d$	Dosage Distribution (DD) for dosage d
$p(\{\mathcal{I}_d\}_{d=1\ldots p-1}\|j)$	Site Dosage Spectrum (SDS) for mutations of frequency j/n
$\psi_{j,\{\mathcal{I}_d\}}$	Site Frequency/Dosage Spectrum (SFDS) for frequency j/n and DD $\{\mathcal{I}_d\}_{d=1\ldots p-1}$
$r_i(x)$	Read depth of the ith individual at position x along the genome
$c_i(x)$	Derived allele count of the ith individual at position x along the genome

our knowledge, all existing statistics for sequences sampled from a single populations at the time of this writing—such as estimators of variability, neutrality tests, estimators based on linkage disequilibrium and haplotype-based statistics—rely implicitly on this assumption.

These statistics can also be classified in terms of the number of sites involved in each individual computation. The frequency of a SNP requires information only on the alleles at a single genomic site, while linkage disequilibrium requires a comparison of alleles at two sites. On the other extreme, haplotype statistics require information on all sites in the sequence.

In this manuscript we will focus on the simplest statistics, i.e., those which can be computed independently for each site (and eventually averaged over all sites in the sequence to obtain summary statistics). We will also consider only biallelic variants (one ancestral and one derived/mutated allele present at each site) in our analysis. Biallelic SNPs represent by far the most common type of variant in eukaryotic genomes, hence this assumption is not particularly restrictive. This is true also for autopolyploid organisms, since it relies on the low mutation rates per base and the corresponding low variability at the population level.

A simple explaination for the prevalence of biallelic variants is the following. Under the usual assumptions for the Kingman coalescent, which describes autopolyploid populations as well (Arnold et al., 2012), SNPs are generated by at least a mutation in a given site along the tree. The tree length in coalescent units is a number of order $O(1)$, while the effective mutation rate in coalescent units is represented by the parameter of genetic variability $\theta = 2pN_e\mu$ where N_e is the effective population size, p is the ploidy and μ is the mutation rate per base. For most eukaryotic organisms, θ is around 10^{-3} (Lynch, 2005). This estimate is based on diploids, but the order of magnitude would be the same for most autopolyploids. The fraction of sites containing a SNP in a finite sample is the product of θ and tree length, and therefore proportional to θ. However, for a triallelic SNP to occur, two mutations should appear on the tree, hence only a fraction $O(\theta^2)$ of sites contains a SNP with three or more alleles, i.e., only a fraction $O(\theta)$ of the SNPs is triallelic. This argument is valid for autopolyploids, but not for allopolyploids, since it does not take into account the divergence between homeologous chromosomes.

In haploid populations, the only statistic based on information at a single position of nucleotide sequences is the frequency of the mutated/derived allele $f(x)$ at a given site x. In fact, once the frequency in the sample is known, the genotypes of all individuals are known up to permutations of the individual. The summary statistic is the so-called SFS, which is the number of sites with a mutation of (derived) frequency j/n in a sample of n individuals, denoted by ξ_j. For the whole population, the equivalent spectrum is the density of sites in the sequence with a mutation of (derived) frequency between f and $f + df$, denoted by $\xi(f)$.

In diploid populations, however, the frequency of a mutation at a given site x is not sufficient to fully determine the genotypes of the n individuals in the sample. The reason is that each individual can be homozygous for either the ancestral or the mutated allele or it can be heterozygous, i.e., it is characterized by an internal count of the mutated allele at that site (which can be 0, 1, or 2) and a corresponding internal frequency (0, 1/2, or 1). Taken together, all individuals in the sample carry an "internal spectrum" distributed as $\mathcal{I}_d(x)$ with $d = 0, 1, 2$, defined as the count of individuals with internal count d for the mutation at position x, which is of course normalized as $\sum_{d=0}^{2} \mathcal{I}_d(x) = n$. This individual spectrum is related to the global frequency of the mutation through its mean count $\sum_{d=0}^{2} d\mathcal{I}_d(x) = 2nf(x)$.

The diploid genotype at position x is fully determined by $\mathcal{I}_d(x)$ up to permutations of the individuals. Given that $\mathcal{I}_d(x)$ has three components (number of ancestral homozygotes $\mathcal{I}_0$, of heterozygotes $\mathcal{I}_1$ and of derived homozygotes $\mathcal{I}_2$) but one is constrained by the number of individuals and another combination corresponds to the frequency, there is only one independent component left, for instance the number of heterozygotes $\mathcal{I}_1(x)$. The information contained in this spectrum is therefore equivalent to the two statistics $f(x)$ and $h(x)$, where $h(x)$ is the *heterozygosity* (the fraction of heterozygous individuals in the sample) defined as $h(x) = \mathcal{I}_1(x)/n$.

Heterozygosity is another very well-known statistic in the population genetics of diploid organisms. If the alleles at site x are in Hardy–Weinberg equilibrium (i.e., under random mating and without selection), the expected fraction of heterozygotes is given by the standard formula $\mathrm{E}[h(x)] = 2f(x)(1 - f(x))$, i.e., it corresponds to the pairwise nucleotide diversity in the population at that site. Its distribution for a discrete sample is a binomial with the same mean $2f(1 - f)$ in terms of the population frequency.

Deviations from the expectation $h \approx 2f(1 - f)$ are signatures of violations of some of the assumptions of the Hardy–Weinberg equilibrium. For example, a deficit of heterozygotes $h < 2f(1-f)$ is expected if there is sub-population structure in the sample, violating the "random mating" assumption.

Note that the most general summary single-site statistic for diploids is neither the SFS nor the heterozygosity, but rather the joint site frequency-heterozygosity spectrum $\psi(f, h)$ or its corresponding version $\psi_{j,\mathcal{I}_1}$ for a finite sample. This joint spectrum is defined as the number of sites with a derived variant at frequency $f = j/2n$ and where a fraction $h = \mathcal{I}_1/n$ of the individuals are heterozygous.

The neutral expectation for this frequency-heterozygosity spectrum in finite samples can be found from the known theory from the frequency spectrum in haploids (Fu, 1995; Ewens, 2004) combined with simple combinatorial arguments applied to the Hardy–Weinberg equilibrium (Weir, 1996). This combination gives

$$\mathrm{E}[\psi_{j,\mathcal{I}_1}] = \frac{\theta\, 2^{\mathcal{I}_1} \frac{n!}{\mathcal{I}_1!\frac{j-\mathcal{I}_1}{2}!\left(n-\frac{j+\mathcal{I}_1}{2}\right)!}}{j\binom{2n}{j}} \tag{1}$$

Note the constraint that $j - \mathcal{I}_1$ should be a multiple of 2.

In **Figure 1**, we illustrate how this spectrum appears under neutrality for a single population of constant size, both in the standard model and under two demographic models: recent admixture and population structure. The latter shows a clear violation of Hardy–Weinberg equilibrium due to a lack of

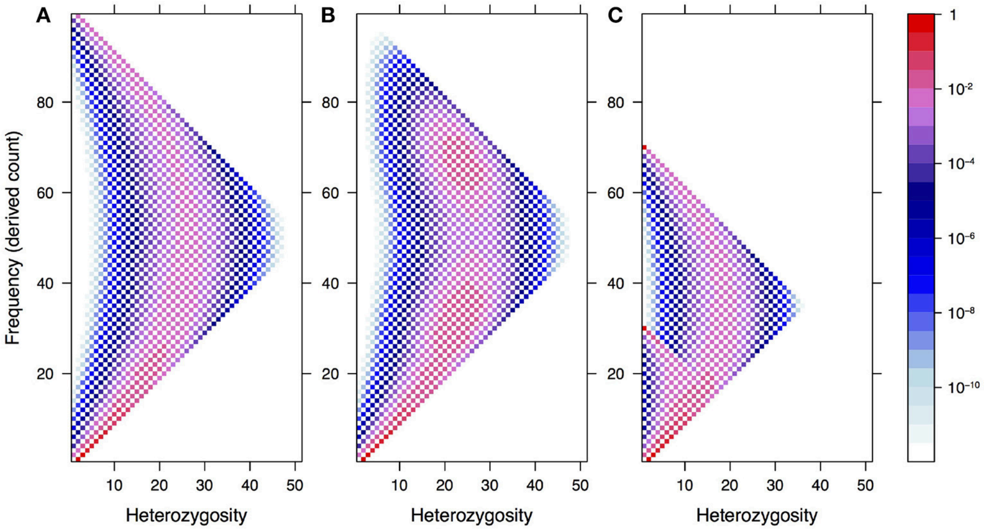


FIGURE 1 | The expected frequency-heterozygosity spectrum for a locus with $\theta = 1$ in a sample of size $n = 100$ from a single population of constant size **(A)** and under two demographic models: recent admixture **(B)** and population structure **(C)**. In both cases, we assume two well-separated populations with divergence equal to θ, the effective population size of the first population being twice the size of the other. In the former case, we assume instantaneous admixture of the two populations and random mating thereafter. In the latter case, the consequence of the absence of mating between different populations is a reduction of heterozygotes in the pooled population, known as the Wahlund effect.

heterozygotes—the so-called Wahlund effect (Rosenberg and Calabrese, 2004).

In diploids, not much attention has been devoted to this joint spectrum, and the two quantities f and h are usually studied separately. One of the possible reasons is that the Hardy–Weinberg equilibrium is reached in a single generation for diploids, hence heterozygosity and deviations from Hardy–Weinberg equilibrium are affected by phenomena acting on short time scales, while the SFS contains information on evolution at larger scales. However, the difference between these quantities becomes more blurred in autopolyploids, as we will discuss in the rest of this paper.

2.2. SFDS in Autopolyploids

In autopolyploids, the framework for single-site statistics is reminiscent of the diploid case. The main difference is that at each position of each individual genome the mutated allele can be present in a number of copies from 0 to the ploidy p. In polyploids, the frequency of an allele within an individual is often called its *allelic dosage*.

The internal spectrum $\mathcal{I}_d(x)$, defined as the count of individuals with allelic dosage d for the mutation at position x, now covers a broader range of dosages $d = 0, 1, 2 \ldots p$. For this reason, we will call it the Dosage Distribution (DD). As before, this spectrum is normalized as $\sum_{d=0}^{p} \mathcal{I}_d(x) = n$ and it is related to the global frequency of the mutation by $\sum_{d=0}^{p} d\mathcal{I}_d(x) = pnf(x)$.

Specification of these two conditions can be avoided if we discard the homozygote counts from the DD, since such counts are completely determined by sample size and frequency together with the rest of the DD. The heterozygous part of the SDS plays the same role as heterozygosity in diploids; however, it has the form of a frequency spectrum, hence an additional complexity with respect to the one-dimensional heterozygosity statistic.

An illustration of the DD and its complexity can be found in **Figure 2**. In this hypothetical example, we consider a panmictic population with mixed mating (partly selfing, partly outcrossing) and distributed according to a spatial density gradient away from a central region. If the selfing rate depends on the density, being low in dense regions and high in sparse ones, then individuals in dense regions will show a pattern consistent with Hardy–Weinberg equilibrium in the DD, while those in sparse regions will show an excess of homozygotes due to selfing.

For large populations, we can define a normalized DD as $i_d = \mathcal{I}_d/n$. The most general single-site statistic for autopolyploids is therefore the joint Site Frequency-Dosage Spectrum (SFDS) $\psi(f, \{i_d\}_{d=1\ldots p-1})$ or its discrete version $\psi_{j,\{\mathcal{I}_d\}_{d=1\ldots p-1}}$ for a finite sample. Similar to the diploid case, this joint SFDS is defined as the number of sites with a derived variant at frequency $f = \frac{j}{pn}$ where the dosage distribution across individuals is $i_d = \mathcal{I}_d/n$. If we condition on a given frequency, we obtain the Site Dosage Spectrum (SDS) $p(\{i_d\}_{d=1\ldots p-1}|f)$.

An important and subtle point that should be clear from **Figure 3** is that the SDS is the distribution of the DD, and hence it cannot be reliably summarized as a single average DD. Reducing the SFDS for a given frequency to the average DD over all variants of that frequency is the equivalent of summarizing the distribution of heterozygosity in diploids by providing the average heterozygosity only. In fact the SFDS is a full p-dimensional spectrum whose components are the frequency (one component) and the heterozygous part of the DD ($p - 1$ components), the latter representing the SDS.

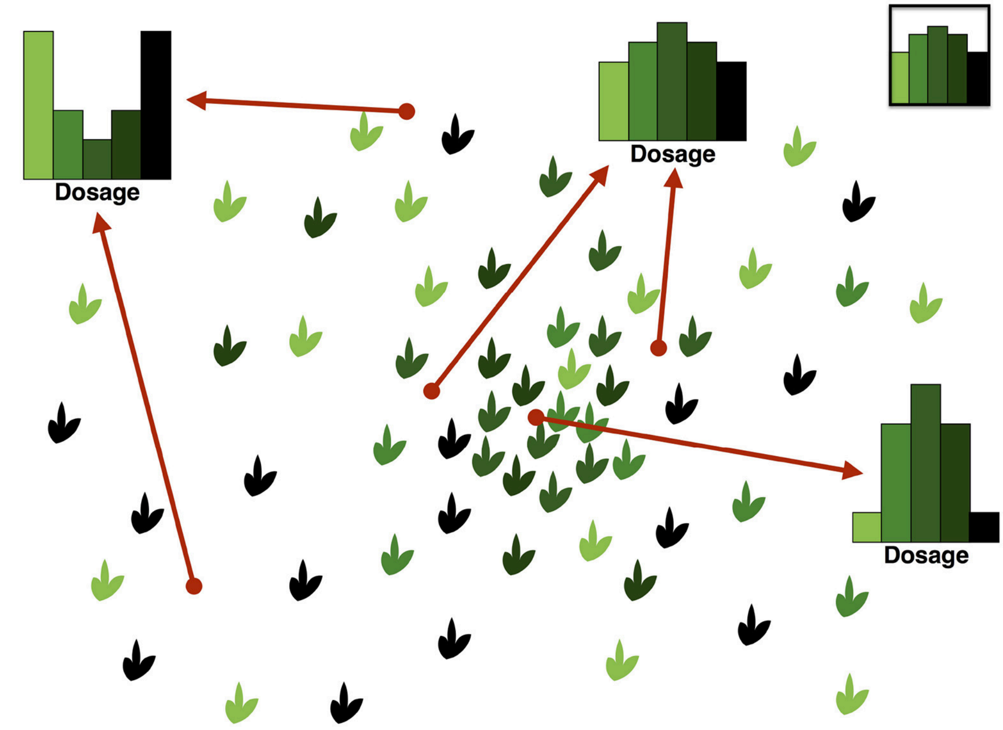


FIGURE 2 | Illustration of the Dosage Distribution (including homozygotes) in a panmictic autotetraploid population with density-dependent selfing rates. In this example, we assume for simplicity that segregating alleles are at intermediate frequency in the population; their dosage in each individual is represented by the color lightness. Since the average frequency is the same everywhere, the average dosage also is. However, by contrast, the DD depends strongly on the sampling location because of variations in the local spatial density. Sampling individuals at random across different locations would result in an average DD like the one in the top-right inset. On the other hand, sampling around a given location would result in different DDs, as illustrated. Locations in the central region tend to have DDs similar to the Hardy–Weinberg ones, while peripheral locations show a large excess of homozygotes because of sampling.

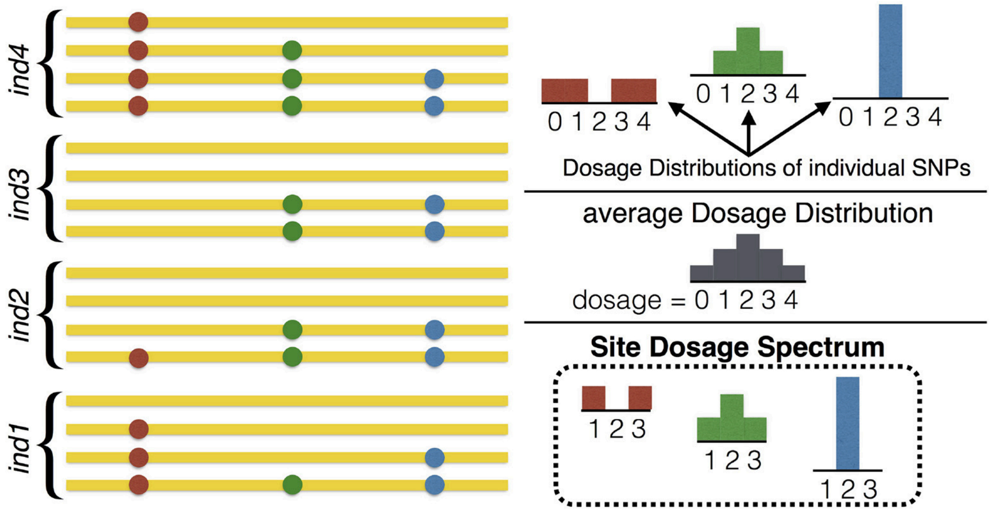


FIGURE 3 | Illustration of the relation between the Dosage Distribution and the Site Dosage Spectrum. On the left, homologous sequences from 4 tetraploid individuals are shown ($n = 4, p = 4$), containing 3 SNPs of frequency 50%. On the right, the three DDs (one for each SNP) are shown at the top. The SDS at the bottom is the distribution of these DDs (which in this example is given by the three DDs with probability 1/3 each). Note that the SDS bears no relation with the average DD, which is shown in the middle. In this example, the Site Frequency/Dosage Spectrum would be $\psi_{8,\{1/4,0,1/4\}} = 1/3$, $\psi_{8,\{1/4,1/2,1/4\}} = 1/3$, $\psi_{8,\{0,1,0\}} = 1/3$ and $\psi_{8,\{\mathcal{I}\}} = 0$ for other choices of $\mathcal{I}$.

2.3. The SFDS of the Standard Neutral Model

The expected value of the SFDS under the standard neutral model is a simple generalization of the diploid frequency-heterozygosity spectrum presented before. In an infinite population and in the absence of double reduction, the Dosage Distribution for a mutation of frequency f under Hardy–Weinberg equilibrium is well-known (Haldane, 1930):

$$i_d = \binom{p}{d} f^d (1-f)^{p-d} \quad \text{for } d = 0 \dots p \tag{2}$$

and the expected value of the neutral SFS has the standard shape

$$\mathrm{E}[\xi(f)] = \frac{\theta}{f}\ ; \tag{3}$$

hence the expected population SFDS is simply

$$\mathrm{E}[\psi(f, \{i_d\})] = \frac{\theta}{f} \prod_{d=1}^{p-1} \delta\left(i_d - \binom{p}{d} f^d (1-f)^{p-d}\right) \tag{4}$$

where $\delta(z)$ is the Dirac delta function, which represents a distribution concentrated at $z = 0$.

For finite samples the expected values are slightly more complex. A combinatorial argument similar to the diploid case — based on the ways to assign the j mutated alleles across the pn homologous chromosomes—provides the following formula for the SDS, i.e., the distribution of the Dosage Distribution $\{\mathcal{I}_d\}_{d=1 \dots p-1}$ in finite samples of size n:

$$\mathrm{E}[p(\{\mathcal{I}_d\}|j)] = \frac{\dfrac{n!}{\mathcal{I}_1!\mathcal{I}_2! \dots \mathcal{I}_{p-1}! \left(\frac{j - \sum_{d=1}^{p-1} d\mathcal{I}_d}{p}\right)! \left(n - \frac{j}{p} - \left(1 - \frac{1}{p}\right)\left(\sum_{d=1}^{p-1} d\mathcal{I}_d\right)\right)!} \displaystyle\prod_{d=1}^{p-1} \binom{p}{d}^{\mathcal{I}_d}}{\dbinom{pn}{j}} \tag{5}$$

where the above expression should be interpreted as 0 if it contains factorials of non-integer numbers. More details can be found in the **Appendix**.

The SFDS in finite samples can be found combining (5) with the known neutral expected SFS θ/j:

$$\mathrm{E}[\psi_{j,\{\mathcal{I}_d\}}] = \frac{\theta}{j} \mathrm{E}[p(\{\mathcal{I}_d\}|j)] \tag{6}$$

Note that in finite samples frequency and DD are under the constraint that $j - \sum_{d=1}^{p-1} d\mathcal{I}_d$ should be a multiple of p.

3. SFS ESTIMATORS AND NEUTRALITY TESTS FOR LARGE SAMPLES

For large samples $n \gg 1$, the exact shape of the DD and the SDS do often have a negligible impact on tests based on the shape of the SFS and their normalization. In fact, most of these tests place weights on $\xi(f)$ that change gradually with the frequency. There are a few exceptions—for instance tests that assign very different weights on singletons, such as Fu and Li's F and D tests for background selection (Fu and Li, 1993), and the expansion test R_2 (Ramos-Onsins and Rozas, 2002). The shape of Hardy–Weinberg violations affects the SFS on a scale $\Delta f \lesssim \frac{p}{pn} = 1/n$. Since most tests weight frequencies in a smooth way over scales of $\Delta f \sim 1/n$ for n large enough, the DD can usually be ignored in large samples.

However, unbiased sequence data from a large number of individuals is typically obtained by High-Throughtput Sequencing (HTS) at low to moderate coverage. HTS data at low coverage is usually unbalanced and more prone to be significantly impacted by sequencing errors, thus requiring tailored approaches. Hence in this section we focus on SFS-based estimators of genetic variability and neutrality tests adapted to HTS data.

SNP calling is usually required prior to population genetic analysis. It is even more relevant for HTS data, due to the typical amount of sequencing errors for these technologies. It is key that only methods developed specifically for polyploids (e.g., GATK from Broad Institute) or for pooled data (e.g., Raineri et al., 2012) are used, since the accuracy of SNP calling algorithms depends on the ploidy. Algorithms for diploids are usually unsuitable to analyse data from organisms with higher ploidy.

Allelic dosage estimation could also be performed (e.g., Blischak et al., 2016), but it is unreliable at low coverage and can be challenging even at high coverage. In fact, dosage uncertainties represent one of the biggest hurdles when dealing with polyploid population genetics (Blischak et al., 2016). However, an accurate estimate of allelic dosage for each individual is not needed to estimate genetic diversity at population level. In fact, none of the methods we discuss in this section requires an explicit estimation of dosage. All these methods work directly on short-read data after SNP calling and filtering of unreliable low-frequency variants.

The estimators of variability proposed in this section take read depth explicitly into account and are unbiased at low coverage as well. Hence there is no need to filter regions of low coverage, although excluding regions with read depth lower than the ploidy could increase the accuracy of the results. However, since our estimators do not take sequencing errors into account, we strongly suggest to perform SNP calling prior to analysing variability with them. For such analyses SNPs can be filtered with moderately conservative parameters, e.g., excluding only SNPs with posterior probability >0.95 or equivalently p-value >0.05 or PHRED quality score <15.

In this section we consider an experimental setup where every polyploid individual of ploidy p in a sample of n individuals is sequenced separately with a read depth of $r_i(x)$ at position x, where $i = 1 \dots n$. The count of the alternative (derived) alleles within reads from the ith individual at position x is $c_i(x)$. If the position x has been filtered out during SNP calling, we discard the SNP and consider $c_i(x) = 0$ for all individuals.

3.1. Estimators of Variability

3.1.1. Watterson's Estimator

The classical estimator of variability based on the SFS is the Watterson estimator (Watterson, 1975), which is based on the number of segregating sites S in a sample of size n. Under an infinite sites model and a panmictic stationary and neutral scenario with population size N, where mutations are randomly and independently occurring given a mutation rate μ per non-overlapped generation (i.e., a Wright-Fisher model), the expected variability level $\theta = 2pN_e\mu$ can be estimated by:

$$\theta_W = \frac{S}{a_n}\,, \tag{7}$$

where $a_n = \sum_{j=1}^{n-1} \frac{1}{j}$. This estimator is based on the expected neutral spectrum of mutations and is sensitive to the presence of an excessive number of singletons (which can be observed, for example, under demographic expansion scenarios (Ramos-Onsins and Rozas, 2002) or in the presence of high rates of artifactual sequencing errors (Achaz, 2008).

A generalization of the Watterson estimator for autopolyploids, in the form of a Maximum Composite Likelihood estimator, has been derived in Equation (34) of Ferretti and Ramos-Onsins (2015). However, this estimator suffers from a strong bias due to sequencing errors. In fact, sequencing errors appear as low frequency variants which increase the estimate of S. Two strategies could be applied to reduce this dependence: either S should be estimated using only filtered SNPs obtained from SNP calling algorithms, or low frequency variants should be removed with an approach similar to that used in Achaz (2008).

3.1.2. Tajima's Estimator of Nucleotide Diversity

Tajima's estimator (Tajima, 1983) or the pairwise nucleotide difference statistic (Π) is also a relevant estimator of nucleotide diversity and is defined as the average number of differences between sequences. In fact, for each position i it estimates the level of heterozygosity in the population [$2f_i(1 - f_i)$, where f_i is the absolute frequency of a given variant allele at position i]. In the infinite-site and stationary neutral model, the expected value of Tajima's estimator (θ_Π) is equal to that of Watterson's estimator (that is, under the ideal Wright-Fisher scenario $\mathrm{E}[\theta_\Pi] = \mathrm{E}[\theta_W] = \theta$). Tajima's estimator for a region of size L is given by:

$$\theta_\Pi = \frac{n}{(n-1)} \sum_{i=1}^{L} 2f_i(1 - f_i). \tag{8}$$

Results from Ferretti et al. (2013) can be combined to build an unbiased estimator of pairwise nucleotide diversity for multiple polyploid individuals:

$$\hat{\theta}_\Pi = \frac{2}{n(n-1)} \left[\frac{p}{p-1} \sum_{j=1}^{n} \pi_j + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \pi_{j,k} \right] \tag{9}$$

where π_j is the average pairwise difference between reads from the jth individual, and $\pi_{j,k}$ is the average pairwise difference between pairs of reads from the jth and kth individual (Ferretti et al., 2013). Both these quantities account naturally for dosage. The factor $p/(p-1)$ is the same factor that appears between the estimates of sample and population heterozygosity in the above formula (8) (Nei and Roychoudhury, 1973).

The above estimator weights the information from all individuals equally, irrespectively of their coverage and dosage. It is possible to build less noisy unbiased estimators by considering further assumptions on the variance of the pairwise differences. Given the average coverage per base $\bar{r}_j$ of the jth individual, the variances can be often approximated by inverse powers of this coverage $\mathrm{Var}(\pi_j) \propto 4/\bar{r}_j + 4/p$, $\mathrm{Var}(\pi_{j,k}) \propto 1/\bar{r}_j + 1/\bar{r}_k + 2/p$ (see **Appendix**). Hence, an approximate Minimum Variance Unbiased Estimator for the pairwise diversity can be obtained by weighting the terms in the above estimator by their variance:

$$\hat{\theta}_\Pi = \frac{\sum_{j=1}^{n} \pi_j \frac{\bar{r}_j(p-1)}{2(\bar{r}_j+p)} + 2\sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \pi_{j,k} \left(\frac{1}{\bar{r}_j} + \frac{1}{\bar{r}_k} + \frac{2}{p} \right)^{-1}}{\sum_{j=1}^{n} \frac{\bar{r}_j(p-1)^2}{2p(\bar{r}_j+p)} + 2\sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \left(\frac{1}{\bar{r}_j} + \frac{1}{\bar{r}_k} + \frac{2}{p} \right)^{-1}} \tag{10}$$

As both versions of this estimator assign a negligible weight to low frequency alleles, they are much more robust with respect to sequencing errors and uncertainties in SNP calling. Hence in the presence of significant rates of sequencing errors, or other related causes of incorrect base calling, any of these estimators should be preferred to the Watterson estimator discussed above.

3.2. Neutrality Tests

3.2.1. Tajima's *D*

Tajima's D test (Tajima, 1989) was the first neutrality test based on the frequency spectrum and it is still the most popular one. It is based on the difference between the Tajima's estimator θ_Π and the Watterson estimator θ_W. As explained above, under the stationary neutral model it is expected that this difference would be zero. However, empirical data violating the theoretical assumptions can result in significant differences. This test can discriminate among some selective and/or demographic processes. The Tajima's D statistic is given by:

$$D = \frac{\hat{\theta}_\Pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}(\hat{\theta}_\Pi - \hat{\theta}_W)}} \tag{11}$$

where the denominator is computed under the standard neutral model and is a function of θ and np.

For HTS data, the numerator of the test can be simply obtained from the difference of the Tajima's and Watterson's estimators presented above.

Obtaining the exact denominator is computationally tricky. A practical approximation is to use the standard denominator for the test, but replacing the "haploid" sample size np by an effective sample size n_{eff} defined as the average number of homologous chromosomes that have been actually sequenced at every position, i.e.,

$$n_{\text{eff}} = \frac{1}{L}\sum_{x=1}^{L}\sum_{j=1}^{n} p\left[1-\left(1-\frac{1}{p}\right)^{r_j(x)}\right] \tag{12}$$

3.2.2. Fay and Wu's *H*

Fay and Wu's H test (Fay and Wu, 2000) was designed to detect derived allele frequencies much higher than expected under a neutral scenario. A large number of variants at high frequencies can be a consequence of positive selection, although it could also occur in the presence of signals of population structure (e.g., introgression). The test compares the levels of variability of Tajima's estimator (θ_Π) vs. another variability estimator—here named θ_H—that weights the number of segregating sites quadratically with the frequency of derived alleles. The normalized version of this test (Zeng et al., 2006) is:

$$H = \frac{\hat{\theta}_\Pi - \hat{\theta}_H}{\sqrt{\text{Var}(\hat{\theta}_\Pi - \hat{\theta}_H)}} \tag{13}$$

For HTS data, we apply the same approach as for Tajima's D. The only difference is that we use the alternative definition of the numerator $2(\theta_\Pi - \theta_L)$ where θ_L is the Zeng's estimator, which is linear in the derived frequency (Zeng et al., 2006). An unbiased version of θ_L for HTS data is

$$\hat{\theta}_L = \sum_{x=1}^{L}\frac{\sum_{j=1}^{n} c_j(x)}{\mathcal{N}_L(x)\sum_{j=1}^{n} r_j(x)} \tag{14}$$

where the normalization factor

$$\mathcal{N}_L = \sum_{k=1}^{pn-1}\frac{1}{k}\sum_{k_1=0}^{p}\cdots\sum_{k_n=0}^{p}\delta_{k,k_1+\ldots+k_n}\frac{\prod_{i=1}^{n}\binom{p}{k_i}}{\binom{pn}{k}}\left[1-\prod_{i=1}^{n}\left(\frac{k_i}{p}\right)^{r_i(x)}\right] \tag{15}$$

is the probability that a segregating site is not interpreted as a fixed derived variant based on the reads. Note that $\delta_{i,j}$ is the Kronecker delta which is 1 if $i = j$ and 0 otherwise.

An approximate version of the denominator of the test can be derived inserting n_{eff} in the standard denominator, as described above for Tajima's D.

4. SMALL SAMPLES AND HARDY–WEINBERG VIOLATIONS IN THE SDS

For small autopolyploid samples, deviations from the neutral SFS cannot be clearly discriminated from violations of Hardy–Weinberg. In fact, in the smallest possible sample of a single individual, the Dosage Distribution coincides with the SFS! More precisely, the SFS for a single individual corresponds to the heterozygous components of the Dosage Distribution averaged across sites. Hence, the features of the DD have a huge impact on the SFS.

This impact is two-fold. On a practical side, if it is not possible to estimate allelic dosage with sufficient accuracy, then uncertainties in individual dosage result in large uncertainties in the determination of allele frequencies, and therefore of the SFS. However in principle, even if dosage could be accurately inferred, the shape of the SFS for a few individuals would still be largely determined by the effect on the DD of the deviations from Hardy–Weinberg equilibrium. We will discuss such deviations in this section.

For diploid organisms there is only one possible direction for Hardy–Weinberg violation, i.e., excess or deficit of heterozygotes. However, in autopolyploids, many different deviations from Hardy–Weinberg equilibria are possible, resulting in different deviations from the neutral SFS. In fact, in this section we present four examples of possible mechanisms of violation of Hardy–Weinberg equilibrium which correspond to four different directions in the space of expected DDs. These examples are (i) inbreeding; (ii) inbreeding with mixed disomic/polysomic inheritance; (iii) heterozygote advantage; (iv) selection against recessive mutations. In tetraploids, combinations of these mechanisms span the whole space of all possible deviations from Hardy–Weinberg.

The shapes of the deviations of the expected DD from a Hardy–Weinberg equilibrium are shown for these mechanisms in **Figure 4**, both in tetraploids and hexaploids. The corresponding directions of the deviations of SFS-based tests from their null values are shown in the same figure for Tajima's D and Fay and Wu's H for a range of ploidy from 4 (tetraploids) to 10 (decaploids).

4.1. Inbreeding

Inbreeding is a well-known cause of violation of Hardy–Weinberg. Both in diploids and in polyploids, selfing and other mechanisms such as subpopulation structure cause a lack of heterozygotes, as discussed in relation to the Wahlund effect (Rosenberg and Calabrese, 2004).

As an example of its consequences on the DD, we can model a small rate of selfing in a population with polysomic inheritance by assuming an equilibrium in the DD given the frequency of the variant, with an approach similar to the one used in De Silva et al. (2005):

$$\mathcal{I}_k^{\text{eq}} = \sum_{k'=0}^{p}\sum_{k''=0}^{p}\mathcal{I}_{k'}^{\text{eq}}\mathcal{I}_{k''}^{\text{eq}}\sum_{a=0}^{p}\text{Hyp}(a|k',p/2,p)\text{Hyp}(k-a|k'',p/2,p) \tag{16}$$

where $\text{Hyp}(\cdot)$ is the hypergeometric distribution that corresponds to the sampling of chromosomes in gametes. Note that all the Hardy–Weinberg equilibrium distributions $\mathcal{I}_k^{\text{eq}} = \binom{p}{k}f^k(1-f)^{p-k}$ discussed before are solutions of the equation above (Here and in the rest of this section, we ignore the possibility of double reduction, since it requires a separate modeling of its impact on allele frequencies; Butruille and Boiteux, 2000).

Then we can perturb the equilibrium by occasional selfing events with a small probability p_s, obtaining:

$$\Delta\mathcal{I}_k = -p_s\mathcal{I}_k^{\text{eq}} + p_s\sum_{k'=0}^{p}\mathcal{I}_{k'}^{\text{eq}}\sum_{a=0}^{p}\text{Hyp}(a|k',p/2,p)\text{Hyp}(k-a|k',p/2,p) \tag{17}$$

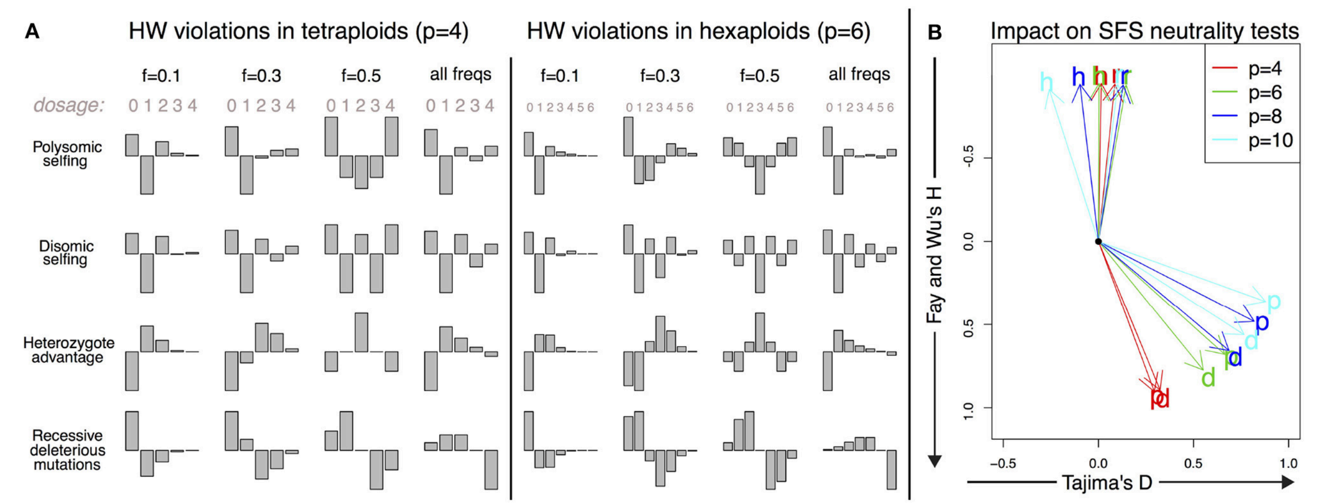


FIGURE 4 | Deviations from the Hardy–Weinberg equilibrium and their impact on the DD. **(A)** Shape of the small deviations $\Delta\mathcal{I}_k$ of the DD from the Hardy–Weinberg equilibrium for both tetraploid and hexaploid individuals in four different scenarios: polysomic selfing (p); disomic selfing (d); heterozygote advantage (h); recessive deleterious mutations (r). We show the deviations for mutations of given frequency (0.1, 0.3, and 0.5) together with the expected violations for random neutral mutations of arbitrary frequency (i.e., distributed as θ/f). The absolute amplitude of the deviations is arbitrarily chosen for each plot; its actual value will depend on parameters such as selfing rates and selection coefficients. **(B)** Impact of the deviations on SFS-based neutrality tests for a single individual. The overall impact is proportional to the amplitude of the deviations; here we show only the directions of apparent violation of neutrality along the space of two SFS-based tests (Tajima's D and Fay and Wu's H). The expected deviations from neutrality are shown for the same four scenarios as in **A** (p, d, h, and r) and for tetraploid, hexaploid, octoploid and decaploid organisms. The black dot corresponds to the neutral values $D = 0$ and $H = 0$.

The shape of this violation of Hardy–Weinberg is shown in **Figure 4**. As expected, it results in an excess of homozygotes in the population. For a single individual, it has a positive impact on both Fay and Wu's H and Tajima's D. For tetraploids, the deviations from the null value are more apparent in H, while in organisms with ploidy higher than 6, violations tend to be larger in D.

4.2. Intermediate Disomic/Polysomic Inheritance

Not only the rates of selfing/outcrossing, but also the mode of inheritance could impact on the violation of Hardy–Weinberg. Mixed disomic/polysomic inheritance is an example of an alternative inheritance mode that appears to be less rare than expected (Meirmans and Van Tienderen, 2013).

Without inbreeding, partial disomic inheritance alone does not lead to violations of the Hardy–Weinberg equilibrium. Hence to study deviations from Hardy–Weinberg we model mixed disomic/polysomic inheritance but with a small selfing rate p_s, similar to the case above. We denote the probability of disomic and polysomic inheritance by p_2 and $1 - p_2$ respectively. For small selfing rate, it is easy to argue that the violations would be a combination of purely disomic and purely polysomic violations with weights p_2 and $1 - p_2$ respectively, i.e.,

$$\Delta\mathcal{I}_k = (1 - p_2)\Delta\mathcal{I}_k^{polysomic} + p_2\Delta\mathcal{I}_k^{disomic} \tag{18}$$

assuming that $p_s \ll 1$.

Purely disomic violations would satisfy similar equations as the purely polysomic ones in the previous section, although with slightly different inheritance terms. Similar to what happens in diploid organisms, sampling of the new generation occurs separately for each heterozygous pair of disomically homologous chromosomes:

$$\Delta\mathcal{I}_k = -p_s\mathcal{I}_k^{\text{eq}} + p_s\sum_{k'=0}^{p}\mathcal{I}_{k'}^{\text{eq}}\sum_{h=0}^{p/2}\frac{2^h\binom{p/2}{h;\,\frac{k'-h}{2};\,\frac{p-k'-h}{2}}}{\binom{p}{k'}}\binom{h}{\frac{k-k'+h}{2}}2^{-h} \tag{19}$$

The corresponding shape of Hardy–Weinberg violations shown in **Figure 4** is similar to the one of selfing in polysomic organisms, but with an excess of homozygous pairs of disomically homologous chromosomes that translates into an excess in the components of even dosage in the spectrum. The impact on Fay and Wu's H and Tajima's D is similar to that of purely polysomic inheritance.

4.3. Heterozygote Advantage

Heterozygote advantage, or overdominance, is a form of "hybrid vigor" where individuals heterozygous for the locus considered acquire a higher fitness than those provided by the two homozygous genotypes. For simplicity, we can assume the two differences in fitness to be the same. Unsurprisingly, this effect tends to increase the amount of intermediate-frequency alleles and heterozygotes (Kaplan et al., 1988).

Modeling selection dependent on the allelic dosage can be done via an approach similar to the one employed above, but is trickier. Selection is not a one-off or rare event but perturbs permanently the equilibrium $\mathcal{I}_k^{\text{eq}}$, hence a self-consistent version of the perturbative equations should be employed. Assigning a fitness $\phi_k = 1 + s_k$ to each allelic dosage, we obtain the equilibrium condition

$$\mathcal{I}_k^{\text{eq}} = \sum_{k'=0}^{p}\sum_{k''=0}^{p} \frac{\mathcal{I}_{k'}^{\text{eq}}\phi_{k'}\mathcal{I}_{k''}^{\text{eq}}\phi_{k''}}{\left(\sum_{l=0}^{p}\mathcal{I}_l^{\text{eq}}\phi_l\right)^2} \sum_{a=0}^{p} \text{Hyp}(a|k', p/2, p)\text{Hyp}(k{-}a|k'', p/2, p) \tag{20}$$

We can then perturb at linear order in s_k and compute $\Delta\mathcal{I}_k = \mathcal{I}_k^{\text{eq}} - \mathcal{I}_k^0$, with $\mathcal{I}_k^0$ being a solution of Equation (16). After using the fact that $\sum_{k=0}^{p}\mathcal{I}_k^0 = 1$, we obtain the linear system

$$\begin{aligned}\Delta\mathcal{I}_k = {} & 2\sum_{k'=0}^{p}\sum_{k''=0}^{p}\mathcal{I}_{k'}^0\left(\mathcal{I}_{k''}^0 s_{k''} + \Delta\mathcal{I}_{k''}\right) \times \\ & \sum_{a=0}^{p}\text{Hyp}(a|k', p/2, p)\text{Hyp}(k-a|k'', p/2, p) \\ & - 2\mathcal{I}_k^0\sum_{l=0}^{p}\left(\mathcal{I}_l^0 s_l + \Delta\mathcal{I}_l\right)\end{aligned} \tag{21}$$

This equation describes how perturbations to the neutral equilibrium driven by weak selection increase, which is a good proxy for the shape of Hardy–Weinberg violations in the DD.

An example of a fitness assignment that leads to heterozygote advantage is $s_k = s$ for $k = 1 \ldots p-1$ but $s_0 = 0$, $s_p = 0$. This gives a constant fitness advantage to all heterozygotes, independently on their dosage.

We report the Hardy–Weinberg violations for this example in **Figure 4**. As expected, heterozygote advantage increases the number of alleles at all frequencies while reducing homozygotes. Surprisingly enough, despite the intuition that the effect would be to increase Tajima's *D* due to the excess of intermediate-frequency variants, the final spectrum impacts negatively on Fay and Wu's *H* and only weakly on Tajima's *D*, as shown in **Figure 4**.

4.4. Recessive Deleterious Mutations

It is possible to use the same approach as in the previous subsection to deal with selection against derived homozygotes. If the mutation is deleterious but recessive, there will be a fitness gap between the homozygotes for the derived allele, which would show the phenotypic effects of the mutation, and all other genotypes, that would not. This is another classical cause of violation of Hardy–Weinberg equilibrium, although in practice it is difficult to detect since the mutations involved tend to be at low frequency and therefore the lack of derived homozygotes could be attributed to the Hardy–Weinberg equilibrium itself.

The fitness assignment for a recessive deleterious allele is $s_p = -s$ but $s_k = 0$ for $k = 0 \ldots p-1$. This describes a selection pressure against derived homozygotes only.

The shape of the Hardy–Weinberg violations in this case shows the expected reduction in derived homozygotes and an excess in intermediate-dosage heterozygotes. This causes a reduction in Fay and Wu's *H*, as shown in **Figure 4**. Ironically, negative values of Fay and Wu's *H* are also one of the typical signatures of selection and genetic hitchhiking.

5. DISCUSSION

In order to advance our understanding of the evolutionary processes affecting the genome of polyploid species, an important step is to gain a deeper knowledge of the way these processes modulate the fate of genetic variants, and consequently the levels and patterns of genetic variability. Two of the main descriptive statistics used in population genetics to summarize genetic variability are the SFS and the heterozygosity (*h*), which contain information on the global and internal allelic spectra, respectively. The expected patterns of these statistics have not been studied in detail for polyploids; that is especially true for many conditions commonly found in empirical studies of autopolyploid species, for instance small sample sizes and violations of the Hardy–Weinberg equilibrium such as inbreeding. In addition, understanding the expected patterns in commonly used statistics such as Tajima's *D* or Fay and Wu's *H* tests is of great relevance for the correct interpretation of the evolutionary processes occurring in autopolyploid populations. Typical patterns there could well be different from the expected patterns in diploid populations, simply because genetic and evolutionary processes have different peculiarities in the two cases.

Studies focused on the analysis of nucleotide variability in polyploid species present special difficulties in comparison to diploid species, as is extensively reviewed in Dufresne et al. (2014). These difficulties have been partially the reason for a relatively scarce number of publications on HTS analysis of genomic variability among wild autopolyploid populations. Nevertheless polyploid plant species in particular are of great interest, given their high economic and strategic impact. In the last years there has been a proliferation of studies on related model species such as *Arabidopsis* (e.g., Hollister et al., 2012; Arnold et al., 2015), other relatively simple species (e.g., Cornille et al., 2016; Kasianov et al., 2017), but also economically important species with more complex genetics (e.g., Raman et al., 2014; Rocher et al., 2015; Kamneva et al., 2017; Krasileva et al., 2017). Although the number of relevant datasets deposited in sequence databases is constantly growing, their adequate analysis will require the further development of specific statistical tools, especially to infer sequence variability and population genomics.

In this manuscript we outlined the rich structure of frequency spectra in autopolyploids. The combination of global and internal spectra—i.e., mutation frequency in the population for the SFS, and allelic dosage in individuals for the SDS—contributes to the complexity of the polyploid SFDS.

The intricacy of the SFS structure and the challenges posed by its correct inference are possibly the reasons why this summary statistic has been given scant attention in polyploids so far (Dufresne et al., 2014; Meirmans et al., 2018), despite the fact that it represents one of the classical statistics in population genetics (Nielsen, 2005; Casillas and Barbadilla, 2017).

In this paper we also discussed some of the challenges related to the analysis of autopolyploid data generated by HTS technologies. However, our discussion is restricted to the simplified case of Hardy–Weinberg equilibrium, which is likely to be violated in many real populations of autopolyploid plants e.g., because of selfing. Even for purely outcrossing autopolyploid organisms, violations of Hardy–Weinberg could be caused by widespread mechanisms such as a large number of recessive deleterious alleles. Similarly, the interplay between the SFS and the Dosage Distribution has been discussed here only in the simplified case of small perturbations of Hardy–Weinberg equilibrium in a single individual. These assumptions allow us to present for the first time a systematic picture of the issues; on the other hand, more work is required to build a theoretical understanding of the SFDS and of SFS-based inference in polyploids, especially for small samples.

One of the most important consequences of the present work is the different interpretation of the neutrality test under deviations from a neutral panmictic model in Hardy–Weinberg equilibrium (**Figure 4**). For a low number of samples, the SFS tends to be dominated by the SDS. Deviations from Hardy–Weinberg equilibrium within each individual distort the full SFS and result in values of neutrality tests that are different from those expected in diploid populations undergoing the same processes. For instance, heterozygote advantage in a small sample of diploid individuals is expected to result in an increase of heterozygotes and therefore a deviation of the Tajima's *D* test toward positive values. On the other hand, in a single autopolyploid individual with the same number of homologous chromosomes, this effect would be close to zero or negative. The reason is two-fold: homozygote alleles would not be classified as polymorphisms and therefore would not be included in the spectrum, while the impact of heterozygote advantage on dosage itself is complex. Generally speaking, the impact of Hardy–Weinberg violations on allelic dosage tends to affect deeply the SFS of the global sample when the sample size is small, complicating the interpretation of the results of neutrality tests. Note that the Hardy–Weinberg equilibrium is not reached in a single generation for autopolyploid species, leaving a longer signal in the genome patterns in relation to diploid species.

The role of allelic dosage uncertainties should be emphasized once more. Despite being challenging, the inference of individual genotypes (i.e., allelic dosage) by likelihood estimation can be obtained from HTS datasets using several algorithms. Recently, Maruki and Lynch (2017) developed a genotype calling algorithm that has proven useful for population genetic analysis. Nevertheless, accurate inference can only be obtained with high read depths and high cost, which usually implies the analysis of just a few individuals. Even in such a case, as shown in this paper, the inference of genotype likelihoods could be hindered by conservative assumptions on the Hardy–Weinberg patterns of the DD, which can generate systematic biases especially in relation to low frequency variants. Focusing on the analysis of variability, the real genotype of each individual is not as important as the pattern of the whole SFS, considering the uncertainties produced by deviations from Hardy–Weinberg equilibrium and other random processes. That is the reason why the equations presented here make performing genotype inference for each autopolyploid individual unnecessary.

Another reason why allelic dosage uncertainty is not a limitation for SFS inference can be illustrated by the following general argument. By definition, the frequency of an allele is the sum of its allelic dosages across individuals divided by the total number of homologous chromosomes in the sample, i.e., np. This implies a relation between frequencies and their uncertainties: more precisely, by classical probability arguments, the standard deviation of the frequency is the quadratic mean of the standard deviation of the allelic dosage divided by $p\sqrt{n}$. Hence, no matter how large is the allelic dosage uncertainty for each individual, the accuracy in the reconstruction of the frequency is always good for samples of large enough size. In fact, the maximum standard deviation of allelic dosage is $p/2$, i.e., the uncertainty in frequency is at most $\frac{1}{2\sqrt{n}}$. This means that 25 individuals are sufficient to estimate allele frequencies with an uncertainty of about 0.1, even in the worst-case estimate of allelic dosage uncertainties.

How large the actual sample should be depends on the actual uncertainties in dosage and the evolutionary dynamics of the population. The typical uncertainties in dosage inference from HTS are expected to be around $p/\sqrt{\bar{r}}$ where $\bar{r}$ is the average read depth per individual, hence they decrease with the sequencing depth of the experiment. However, if the dynamics is driven by rare variants, a larger number of individuals is needed to obtain an accurate estimate of their frequency, since the unavoidable variance in frequency due to the sampling process of individuals from the whole population is between $\frac{f(1-f)}{pn}$ (under Hardy–Weinberg equilibrium) and $\frac{f(1-f)}{n}$ (if the Hardy–Weinberg conditions are strongly violated).

At present, the complexity of most analyses implies that good-quality population genetic data of samples of multiple autopolyploid organisms from the same natural population are hard to obtain. Most of the efforts so far were focused on the relation between different populations (Meirmans and Hedrick, 2011) and the comparison between different levels of ploidy, which require the sequencing of single samples from multiple populations. On a broader evolutionary scale, polyploidization during speciation and its evolutionary consequences were also studied in several biological systems (Parisod et al., 2010; Barker et al., 2016). However, there is a general lack of good datasets, and theoretical approaches to understand the microevolutionary picture are lagging behind (Dufresne et al., 2014; Meirmans et al., 2018), with the possible exception of linkage and QTL mapping. We hope that this paper will raise some awareness of the issues involved and clarify the relation between important quantities such as the frequency spectrum, the heterozygosity and the distribution of allelic dosage.

In conclusion, considering spectra of allelic dosage such as the SDS is of fundamental importance for the study of the evolutionary processes in autopolyploids. These internal spectra have a large impact on the global SFS for small sample sizes (for large sample size, the SFS can be reliably inferred and should not be strongly affected by Hardy–Weinberg violations). In this framework, we have proposed a set of estimators of variability and neutrality tests for autopolyploid HTS samples, based on well-known tests such as Tajima's *D* and Fay and

Wu's *H*. Additionally, we have shown how different deviations from Hardy–Weinberg equilibrium and other uncertainties are reflected in the dosage distribution at the level of single individuals. In general, we bring attention to the importance of the study of the joint SFDS in polyploid species in order to correctly interpret the patterns of population variability.

AUTHOR CONTRIBUTIONS

LF and SR-O conceived the paper. LF and PR developed the theory. LF implemented it. LF, PR, and SR-O wrote the paper.

FUNDING

This work was supported by grant AGL2016-78709-R (MEC, Spain) to SR-O. We also acknowledge the financial support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016-2019 (SEV-2015-0533) grant awarded to the Center for Research in Agricultural Genomics and by the CERCA Programme/Generalitat de Catalunya. The Pirbright Institute receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom (projects BB/E/I/00007035, BB/E/I/00007036 and BBS/E/I/00007039).

ACKNOWLEDGMENTS

We thank Emanuele Raineri and Miguel Pérez-Enciso for past discussions on the development of individual HTS estimators. We also thank the editors of this Special Issue on Polyploid Population Genetics and Evolution for their suggestion to submit this paper. Further details on the mathematical derivations and the R code used to generate the figures can be found in **Supplementary Material**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00480/full#supplementary-material

REFERENCES

Achaz, G. (2008). Testing for neutrality in samples with sequencing errors. *Genetics* 179, 1409–1424. doi: 10.1534/genetics.107.082198

Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183, 249–258. doi: 10.1534/genetics.109.104042

Arnold, B., Bomblies, K., and Wakeley, J. (2012). Extending coalescent theory to autotetraploids. *Genetics* 192, 195–204. doi: 10.1534/genetics.112.140582

Arnold, B., Kim, S.-T., and Bomblies, K. (2015). Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Mol. Biol. Evol.* 32, 1382–1395. doi: 10.1093/molbev/msv089

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Blischak, P. D., Kubatko, L. S., and Wolfe, A. D. (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol. Ecol. Resour.* 16, 742–754. doi: 10.1111/1755-0998.12493

Butruille, D. V., and Boiteux, L. S. (2000). Selection-mutation balance in polysomic tetraploids: impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6608–6613. doi: 10.1073/pnas.100101097

Casillas, S., and Barbadilla, A. (2017). Molecular population genetics. *Genetics* 205, 1003–1035. doi: 10.1534/genetics.116.196493

Chester, M., Gallagher, J. P., Symonds, V. V., Cruz da Silva, A. V., Mavrodiev, E. V., Leitch, A. R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 109, 1176–1181. doi: 10.1073/pnas.1112041109

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., et al. (2016). Genomic signature of successful colonization of eurasia by the allopolyploid shepherd's purse (capsella bursa-pastoris). *Mol. Ecol.* 25, 616–629. doi: 10.1111/mec.13491

De Silva, H. N., Hall, A. J., Rikkerink, E., McNeilage, M. A., and Fraser, L. G. (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity* 95, 327–334. doi: 10.1038/sj.hdy.6800728

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Ewens, W. J. (2004). *Mathematical Population Genetics. I. Theoretical Introduction. Interdisciplinary Applied Mathematics*. New York, NY: Springer-Verlag. doi: 10.1007/978-0-387-21822-9

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive darwinian selection. *Genetics* 155, 1405–1413.

Ferretti, L., Perez-Enciso, M., and Ramos-Onsins, S. (2010). Optimal neutrality tests based on the frequency spectrum. *Genetics* 186, 353–365. doi: 10.1534/genetics.110.118570

Ferretti, L., Raineri, E., and Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. *Genetics* 191, 1397–1401. doi: 10.1534/genetics.112.139949

Ferretti, L., and Ramos-Onsins, S. E. (2015). A generalized watterson estimator for next-generation sequencing: from trios to autopolyploids. *Theor. Popul. Biol.* 100C, 79–87. doi: 10.1016/j.tpb.2015.01.001

Ferretti, L., Ramos-Onsins, S. E., and Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Mol. Ecol.* 22, 5561–5576. doi: 10.1111/mec.12522

Fu, Y. X. (1995). Statistical properties of segregating sites. *Theor. Popul. Biol.* 48, 172–197. doi: 10.1006/tpbi.1995.1025

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Gao, H., Williamson, S., and Bustamante, C. D. (2007). A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176, 1635–1651. doi: 10.1534/genetics.107.072371

Haldane, J. B. S. (1930). Theoretical genetics of autopolyploids. *J. Genet.* 22, 359–372. doi: 10.1007/BF02984197

Hardy, O. J. (2016). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Mol. Ecol. Resour.* 16, 103–117. doi: 10.1111/1755-0998.12431

Hollister, J. D., Arnold, B. J., Svedin, E., Xue, K. S., Dilkes, B. P., and Bomblies, K. (2012). Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* 8:e1003093. doi: 10.1371/journal.pgen.1003093

Jost, L. (2008). G(st) and its relatives do not measure differentiation. *Mol. Ecol.* 17, 4015–4026. doi: 10.1111/j.1365-294X.2008.03887.x

Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (fragaria) using haplotypes generated from target capture sequencing. *BMC Evol. Biol.* 17:180. doi: 10.1186/s12862-017-1019-7

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics* 120, 819–829.

Kasianov, A. S., Klepikova, A. V., Kulakovskiy, I. V., Gerasimov, E. S., Fedotova, A. V., Besedina, E. G., et al. (2017). High-quality genome assembly of capsella bursa-pastoris reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J.* 91, 278–291. doi: 10.1111/tpj.13563

Kingman, J. (1982). The coalescent. *Stochastic Process. Appl.* 13, 235–248. doi: 10.1016/0304-4149(82)90011-4

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., et al. (2017). Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 114, E913–E921. doi: 10.1073/pnas.1619268114

Lynch, M. (2005). The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 23, 450–468. doi: 10.1093/molbev/msj050

Maruki, T., and Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3* 7, 1393–1404. doi: 10.1534/g3.117.039008

Meirmans, P. G., and Hedrick, P. W. (2011). Assessing population structure: F(st) and related measures. *Mol. Ecol. Resour.* 11, 5–18. doi: 10.1111/j.1755-0998.2010.02927.x

Meirmans, P. G., Liu, S., and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Hered.* 109, 283–296. doi: 10.1093/jhered/esy006

Meirmans, P. G., and Van Tienderen, P. H. (2013). The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* 110, 131–137. doi: 10.1038/hdy.2012.80

Mielczarek, M., and Szyda, J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *J. Appl. Genet.* 57, 71–79. doi: 10.1007/s13353-015-0292-7

Nei, M., and Roychoudhury, A. K. (1973). Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.* 107, 362–372. doi: 10.1086/282840

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154, 931–942.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.* 39, 197–218. doi: 10.1146/annurev.genet.39.073003.112420

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., and Perez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics* 13:239. doi: 10.1186/1471-2105-13-239

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., et al. (2014). Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. *PLoS ONE* 9:e101673. doi: 10.1371/journal.pone.0101673

Ramos-Onsins, S. E. and Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* 19, 2092–3100. doi: 10.1093/oxfordjournals.molbev.a004034

Rocher, S., Jean, M., Castonguay, Y., and Belzile, F. (2015). Validation of genotyping-by-sequencing analysis in populations of tetraploid alfalfa by 454 sequencing. *PLoS ONE* 10:e0131918. doi: 10.1371/journal.pone.0131918

Rosenberg, N. A., and Calabrese, P. P. (2004). Polyploid and multilocus extensions of the wahlund inequality. *Theor. Popul. Biol.* 66, 381–391. doi: 10.1016/j.tpb.2004.07.001

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). Dna sequencing at 40: past, present and future. *Nature* 550, 345–353. doi: 10.1038/nature24286

Stift, M., Berenos, C., Kuperus, P., and van Tienderen, P. H. (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to rorippa (yellow cress) microsatellite data. *Genetics* 179, 2113–2123. doi: 10.1534/genetics.107.085027

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics* 105, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276. doi: 10.1016/0040-5809(75)90020-9

Weir, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates.

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* 9:104. doi: 10.3389/fpls.2018.00104

Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174, 1431–1439. doi: 10.1534/genetics.106.061432

Conflict of Interest Statement: 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.

METHODS
published: 13 April 2018
doi: 10.3389/fgene.2018.00123

Inferring Variation in Copy Number Using High Throughput Sequencing Data in R

*Brian J. Knaus and Niklaus J. Grünwald**

Horticultural Crops Research Unit, United States Department of Agriculture-Agricultural Research Service, Corvallis, OR, United States

Inference of copy number variation presents a technical challenge because variant callers typically require the copy number of a genome or genomic region to be known *a priori*. Here we present a method to infer copy number that uses variant call format (VCF) data as input and is implemented in the R package *vcfR*. This method is based on the relative frequency of each allele (in both genic and non-genic regions) sequenced at heterozygous positions throughout a genome. These heterozygous positions are summarized by using arbitrarily sized windows of heterozygous positions, binning the allele frequencies, and selecting the bin with the greatest abundance of positions. This provides a non-parametric summary of the frequency that alleles were sequenced at. The method is applicable to organisms that have reference genomes that consist of full chromosomes or sub-chromosomal contigs. In contrast to other software designed to detect copy number variation, our method does not rely on an assumption of base ploidy, but instead infers it. We validated these approaches with the model system of *Saccharomyces cerevisiae* and applied it to the oomycete *Phytophthora infestans*, both known to vary in copy number. This functionality has been incorporated into the current release of the R package *vcfR* to provide modular and flexible methods to investigate copy number variation in genomic projects.

Keywords: bioinformatics, computational biology, copy number variation (CNV), high throughput sequencing (HTS), *Phytophthora*, ploidy, R package

OPEN ACCESS

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Beniamino Trombetta,
Sapienza Università di Roma, Italy
Garrett McKinney,
University of Washington,
United States

****Correspondence:***
Niklaus J. Grünwald
nik.grunwald@ars.usda.gov

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

Received: *31 January 2018*
Accepted: *26 March 2018*
Published: *13 April 2018*

Citation:
Knaus BJ and Grünwald NJ (2018) Inferring Variation in Copy Number Using High Throughput Sequencing Data in R. Front. Genet. 9:123. doi: 10.3389/fgene.2018.00123

INTRODUCTION

Investigations into the variation in the number of copies of genes, chromosomes, or genomes are well-established research topics, yet they continue to present technical challenges to molecular genetic analysis. Many examples provide evidence of how copy number affects the phenotype. For example, schizophrenia in humans is thought to be caused by variation in copy number of certain genes (Sekar et al., 2016). Presence of an additional chromosome (aneuploidy) results in Down syndrome in humans (Hassold and Hunt, 2001). Existence of an extra copy of all chromosomes (triploidy) is used in agriculture to produce sterile organisms such as seedless watermelons (Varoquaux et al., 2000) or sterile salmon (Johnstone, 1992; Cotter et al., 2000). Whole genome duplication (polyploidy) results in every chromosome being duplicated, a phenomenon observed throughout plants, animals, and fungi (Todd et al., 2017; Van de Peer et al., 2017). Although this phenomenon is well established, it presents a challenge to high throughput sequencing projects in that most popular genomic variant callers, such as the GATK's (DePristo et al., 2011) or FreeBayes (Garrison and Marth, 2012), require the *a priori* specification of how many alleles to call. While the inference of copy number may be an important precursor to point mutation discovery, many authors argue that copy number variation may be more abundant throughout a genome than point mutations (Katju and Bergthorsson, 2013) making it an important facet in the investigation of genomic architectures.

Existing software for determining the number of copies at a locus from high throughput sequencing data can be broadly classified into two categories: copy number variation detection and whole genome ploidy inference. The important difference among these categories is the form of data they use. Copy number variation detection software uses per position sequence depth (Yoon et al., 2009; Abyzov et al., 2011; Klambauer et al., 2012; Li et al., 2012) while whole genome ploidy inference software uses the relative frequency of the two most abundant alleles sequenced at a locus (Zohren et al., 2016; Gompert and Mock, 2017; Weiß et al., 2018). Copy number variation detection methods group the per position sequence depth into windows and attempt to sort these into base-ploid (typical depth) windows or windows that deviate from base-ploid. They generally require the investigator to specify *a priori* what copy level the base-ploid state is. If the research question is to determine how many copies occur at the base-ploid state, these methods will not be appropriate. Whole genome ploidy inference methods use the frequency that the two most abundant alleles were sequenced at for heterozygous positions, or allele balance, and summarize this information throughout the genome. (Here we use the term 'allele balance' where other authors have used 'allele frequency' to distinguish the measure from the use of 'allele frequency' in population genetics.) For example, for heterozygous alleles we would expect to observe an approximate frequency of one half for diploids, ratios of thirds for triploids, and ratios of quarters for tetraploids (**Figure 1**). Whole genome ploidy inference uses all of the genomic information to infer a single copy number for the entire genome. A third hybrid method uses allele balance (referred to as allelic ratio) and heterozygosity to assign copy number to populations of data (McKinney et al., 2017). However, if the research question is to explore copy number variation within a population this method will not be relevant. Therefore, there are at least two distinct approaches to determine the number of copies present in genomes, and more currently being proposed, each with different strengths and limitations.

Our research presented us with the need to determine if copy number varied throughout genomes, where we did not have prior knowledge of what the actual base-ploidy might be. We therefore combined the windowing functionality from copy number variation detection methods with the allele balance concept from whole genome ploidy inference methods. We use a non-parametric approach to infer copy number given that empirical explorations of available data indicated that common distributions, particularly at low sequence depth, do not fit well. Our method is implemented in a new update to the package *vcfR* in the R software environment (R Core Team, 2018). R is an established and growing language facilitating the analysis of population genetic and genomic data (Paradis et al., 2017a,b). We demonstrate the utility of this method using genomes from the model fungus *Saccharomyces cerevisiae* and our ongoing work with the oomycete plant pathogen *Phytophthora infestans*. Both of these organisms show variation in ploidy across individuals as well as within regions within a genome.

MATERIALS AND METHODS

Methodology

We developed new functionality added to the current release of the *vcfR* package that can be used to infer copy number or ploidy in R. We initially developed *vcfR* for VCF data import/export,

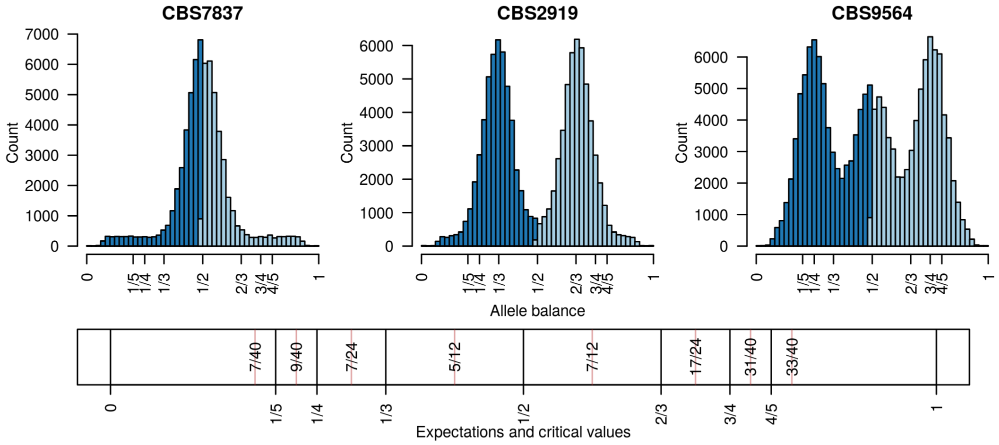


FIGURE 1 | Allele balance (e.g., the distribution of the frequency at which the most abundant allele and the second most abundant allele were sequenced) at heterozygous positions in three *Saccharomyces cerevisiae* genomes. For each heterozygous genotype the frequency at which the most abundant allele was sequenced at (light blue) and the frequency at which the second most abundant allele was sequenced at (dark blue) were recorded. This information was then summarized with a histogram. Expectations for allele balance are 1/2 for diploids, 1/3 and 2/3 for triploids, and 1/4, 1/2, and 3/4 for tetraploids. This approach provides a dominant copy number for each genome but no information about variation within each genome. Expectations and critical values for binning allele balance information are presented below the histograms.

quality control, visualization and general manipulation (Knaus and Grünwald, 2017). *vcfR* now includes a range of new functions useful for binning variants into windows, summarizing the frequency that alleles were sequenced at, and assigning a closest expected copy number value to these windows (**Table 1**).

Data from high throughput sequencing (HTS) projects on populations typically results in calling variants that might include single nucleotide polymorphisms (SNPs), indels, and inversions. Output from popular variant callers is presented in files that adhere to the variant call format (VCF) specification (Danecek et al., 2011). This specification provides the option to include counts for how many times each allele was sequenced for each genotype. For example, in the GATK's HaplotypeCaller (McKenna et al., 2010) output includes allele depth (AD) as a comma delimited string of counts. This VCF data can be imported into R using our function read.vcfR(). Once any desired quality control steps have been performed on the data (Knaus and Grünwald, 2017), such as omitting variants of unusual sequence coverage, this allele depth data can be extracted using the *vcfR* function extract.gt(). We then use the function is_het() to set homozygous positions in the allele depth matrices as missing data (NA) so we can focus our analysis on the heterozygous positions. The allele depth is reported as a comma delimited string, the individual elements of which can be isolated with the function masplit(). Dividing the count for each allele by the sum of the counts for the two most abundant alleles, results in the frequency at which each allele was sequenced, or allele balance. This data can now be plotted as histograms for visualization.

Determining copy number for sub-genomic regions requires the genome to be divided into sub-genomic windows and, because this typically results in many windows per sample, it requires a numeric method of summarizing this data. This goal is accomplished with the function freq_peak(). This function takes as input a matrix of allele balance data, as described above, a vector of chromosomal positions for each variant, a window size, and a bin width for summarizing the allele balance values. The vector of chromosomal positions is used to assign variants to windows. The window size specifies how large the genomic windows should be. This will in part be based on the frequency of heterozygous positions observed in the target sample as well as a balance between the conflicting desires for small windows that provide fine scale resolution and large windows that provide a large number of variants (i.e., support) for a determination. Within each window the allele balance values are summarized by bins from 0 to 1 and of the width specified by the bin width parameter. The bin with the greatest number of variants is selected as the peak location. Here, again, a balance must be found between resolution (small bins) and support (large bins). Default values are provided based on what we have determined to work in our study systems, but we highly encourage adjusting the parameters based on the specifics of each project. These parameters are expected to be context specific to each study system. This function returns three matrices, one containing the window coordinates, one containing the peak locations and one containing the count of variants that resulted for each window. The matrix of variant counts per window can be used to help determine optimal window size and to censor windows that resulted in a low number of variants. The peaks can then be assigned to their nearest expected value (1/5, 1/4, 1/3, 1/2, 2/3, 3/4, 4/5) using the function peak_to_ploid(). This is accomplished by using critical values that are half way between each expected value (**Figure 1**). Once a copy number has been assigned its confidence is measured by creating a distance from expectation. The distance from expectation is the observed value subtracted by the expectation it was assigned to which is then divided by the critical value on the side of the expectation where the observed value was (**Figure 1**). Dividing the critical value scales the difference from expectation from zero (exactly at our expectation) to one (half way between expectations). This can also be used to remove border cases where observed value is intermediate to the expected values and we therefore have low confidence in the determination. The results from the function freq_peak() can be visualized using freq_peak_plot(). This last function was inspired in part by BAF plots (Laurie et al., 2010).

Theoretical population genetics is based largely on haploid and diploid organisms. Investigations into populations that consist of higher ploidy individuals, or populations with a mixture of copy numbers, present a methodological challenge in that few applications are available to analyze them. We have extended Nei's G_{ST} (Nei, 1973, 1987) and Hedrick's G'_{ST} (Hedrick, 2005) to address this challenge. These measures of population subdivision are based on ratios of heterozygosity. Because heterozygosity is based on the number and type of alleles found in a population it provides a convenient way to analyze populations of mixed copy number. Our implementation is inspired by the implementation in *adegenet* (Jombart, 2008) which weights the heterozygosities by their sample size. This is an attempt to correct for unbalanced sample sizes, situations where a different number of individuals were sampled from different populations. We instead weight the heterozygosities by the observed number of alleles in each population to correct for both unbalanced samples as well as instances where individuals may vary in copy number as well. An unbalanced design occurs when different amounts of data are collected for different populations. For example, one sample may have consisted of 20 individuals while another may have only consisted of 10. This imbalance may

TABLE 1 | Functions available to analyze copy number variation and mixed copy number data in the current release of *vcfR*.

Function	Description
extract.gt()	Isolate data from the delimited VCF genotype fields.
freq_peak()	Windowize and identify peaks of density.
is_het()	Identify heterozygous variants.
masplit()	Isolate values from a matrix of delimited data.
peak_to_ploid()	Convert peaks of density to an expected copy number.
freq_peak_plot()	Visualize results from freq_peak().
rePOS()	Convert chromosomal positions to genomic (non-overlapping) positions.
genetic_diff()	Calculate genetic differentiation (G_{ST}).

have occurred due to logistical reasons or technical issues in sample preparation. When copy number is unknown, the investigator may sample the same number of individuals in the populations, but if one population turns out to have four copies where the other has only two, the population with four copies will have twice as much information as the other. Weighting each population by the number of alleles observed is an attempt to mitigate these issues. The function genetic_diff() uses a *vcfR* object and a factor that indicates population membership (VCF data typically does not include population information) and returns a table including heterozygosities, Nei's G_{ST}, and Hedrick's G'_{ST}.

Example Data

To demonstrate our method, we tested it on three data sets. The first data set consisted of three samples of *Saccharomyces cerevisiae* (CBS7837, CBS2919, and CBS9564) from Zhu et al. (2016) that were reported as diploid, triploid and tetraploid by Weiß et al. (2018). We also included an additional sample (YJM1098) that was reported by Zhu et al. (2016) as being predominantly diploid but demonstrating aneuploidy for chromosome XII. These samples represent an organismal system where the genome is of relatively small size (12 Mbp), high quality (in its 64th revision; Engel et al., 2014) and where the samples were sequenced with a goal of attaining 80X sequence depth with Illumina GAII reads.

A second data set consisted of two samples of the plant pathogen *Phytophthora infestans* (99189 and 88069) that were reported by Weiß et al. (2018) as being diploid and triploid. The *P. infestans* system represents a more modestly sized genome (240 Mbp) that remains in its first draft (Haas et al., 2009), but where the samples were sequenced with the intent of attaining 100X sequence depth for each haplotype using Illumina HiSeq 3000 sequencing (Weiß et al., 2018).

The third dataset included 17 samples of *P. infestans* and one sample of *P. mirabilis* collected from the literature, subset to Supercontig_1.50, and made available as an R package (Knaus and Grünwald, 2017). This represents a set of samples that were of more typical sequence depth for genomics projects than we might expect from investigations that were specifically interested in copy number.

For the first two datasets, the data were downloaded from the NCBI sequence read archive and FASTQ data were extracted using the sratoolkit. These reads were mapped to the yeast genome (S288C) or the *P. infestans* genome (T30-4) using bwa 0.7.10-r789 mem (Li, 2013). The resulting SAM file had mate pair information updated, was sorted and converted to BAM format using samtools 1.3.1 (Li et al., 2009). Duplicates were marked using picard-tools-2.5.0 and the files were indexed using samtools. For each sample, a g.VCF file was created from its BAM file using the GATK's (3.5-0-g36282e4) HaplotypeCaller (McKenna et al., 2010). Read processing for the pinfsc50 was described previously (Knaus and Grünwald, 2017). Briefly, the reads were mapped using bwa mem and variants were called using the GATK's HaplotypeCaller resulting in VCF data. The g.VCF and VCF data were processed in *vcfR* (Knaus and Grünwald, 2017) using the methods described above using the functions freq_peak(), peak_to_ploid(), and freq_peak_plot(). For the *S. cerevisiae* samples, a window size 40 kbp was used while a window size of 200 kbp was used for the *P. infestans* samples.

Performance

We assessed performance of our method over a range of genome sizes. Data used for the benchmarking were subset from the 99189 *P. infestans* sample including the entire data set (240 Mbp genome) and subsets of this dataset to represent genomes of 100, 10, and 1 Mbp. Each data set was processed 20 times and this processing was implemented using an R markdown script. The use of R markdown, as opposed to a pure R script, likely incurred a performance cost as our timing included the compilation of the R markdown to a web page. We advocate that using tools like R markdown should be considered a best practice and hope that this will characterize typical use. Benchmarking was performed on an Intel© Core^{TM} i7-4790 CPU at 3.60 GHz with 32 GB of RAM running Ubuntu 16.04 LTS. Results were visualized in R and a linear regression was performed using the R function stats::lm().

RESULTS

Implementation

A new update for the R package *vcfR* was recently released including several new functions (**Table 1**). The function freq_peak() returns the peaks called for each window as well as diagnostic information. The data in VCF files only includes information for the variable positions. This means that all positions in a window will not be present in VCF data. A lookup table is created and returned that includes the genomic coordinates for each window, the row number of the first and last rows of VCF data that were analyzed, and the genomic position of the first and last variant in each window. This information is intended to coordinate comparisons among data extracted from VCF files and genomic windows. A matrix of variant counts per sample and window is also provided. Because heterozygosity may not be known and some windows may have mapping issues (e.g., high variant counts) or regions of loss of heterozygosity or a high number of missing or ambiguous nucleotides in the reference (low variant counts), this information can be used to help determine optimal window size for a particular organism. Furthermore, this approach can help identify anomalous regions in the genome that may require further scrutiny. Lastly, a matrix of frequencies of allele balance is generated.

Results of the above process can be visualized and post-processed to obtain copy number calls and quality assessment. The function freq_peak_plot() can be used to visualize the combined VCF derived data and the results of the windowing and peak calling operations. Because the result is a simple data structure (a list of matrices) the universe of R packages that can be used with matrix data are also available to explore the data. The data can also be post processed with the function peak_to_ploid()

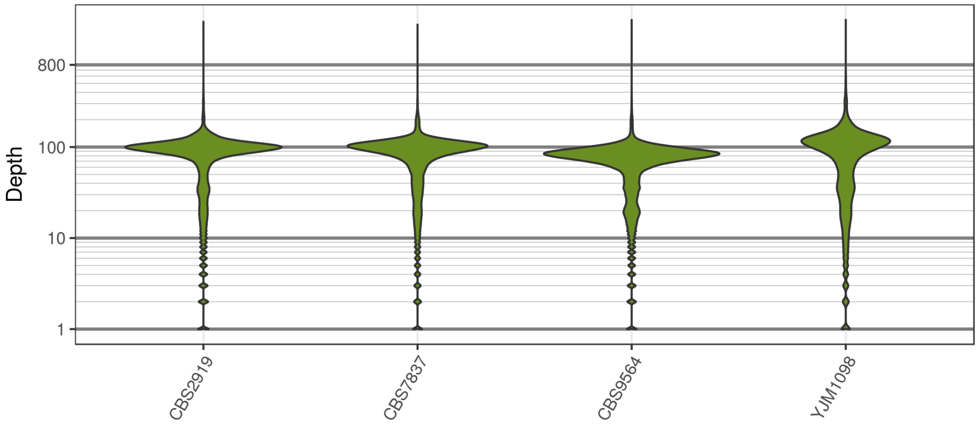


FIGURE 2 | The distribution of sequence depth at variable positions in *Saccharomyces cerevisiae*. While each genome was sequenced at close to 100X, each genome also had long tails for variants that were sequenced at very high and low coverage. These tails are typically observed for high throughput sequencing data.

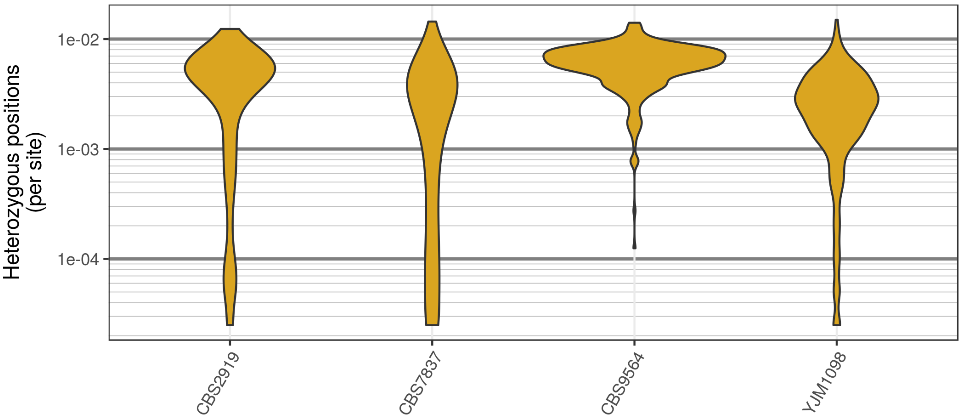


FIGURE 3 | Genomic distribution of heterozygous positions in *Saccharomyces cerevisiae* genomes. Each genome was divided into 40 kbp windows, the number of variants was counted within each window, and this count was divided by the window size. While most windows had a typical number of heterozygous positions (2–8 per kbp) there were a substantial number of windows that contained very few heterozygous positions. Note that these are raw variants from the VCF file produced by the variant caller (in our case, GATK HaplotypeCaller). Because most variant callers take an aggressive perspective on variant calling, the values presented are likely an over-estimate of heterozygosity.

that converts the allele balance frequency data to an integer copy number as well as distances from expectation:

$$\text{Distance from expectation} = \frac{\text{observed allele balance} - \text{expected value}}{\text{critical value}}$$

The distance from expectation is the observed allele balance frequency subtracted by the frequency expected based on the final determination. This value is then divided by its bin width (**Figure 1**) in order to scale it from zero to one where zero represents an allele balance that is exactly on our expectation (e.g., 1/4, 1/3, 1/2, etc.) and one is half way between two expectations. This value can then be used as a measure of confidence in our copy number determination and to omit border cases (instances where the observed allele balance is close to one).

Saccharomyces cerevisiae Dataset

Analysis of the *Saccharomyces cerevisiae* dataset validated previous reports and revealed new features. The *S. cerevisiae* samples were sequenced at about 100X at variable positions (**Figure 2**) making it a high coverage dataset. The samples were determined to consist of individuals that were predominantly diploid (CBS7837), triploid (CBS2919), and tetraploid (CBS9564), confirming previous reports (**Figure 1**; Weiß et al., 2018). The samples had a heterozygosity of around 0.003–0.008 heterozygous positions per site (**Figure 3**). Because the variant caller (the GATK's HaplotypeCaller) tends to aggressively call variants, this estimate may include false positives and therefore may be an overestimate of the true biological value. We have previously discussed strategies we feel may improve the quality of called variants to attain a production data set (Knaus and Grünwald, 2017). Current functionality in *vcfR* allowed for convenient reproduction of

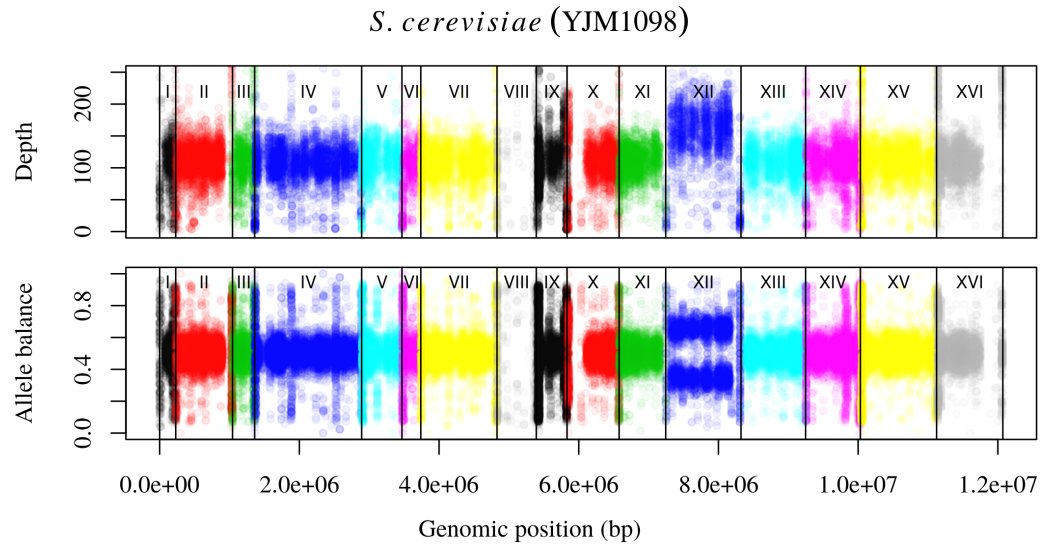


FIGURE 4 | Reproduction of Figure 7 from Zhu et al. (2016). The upper panel demonstrates the concept of base ploidy where most of the genome is of one ploidy however, we do not know how many copies this base ploidy consists of. The lower panel demonstrates how allele balance is predominantly what we would expect for a diploid, allowing us to assign a copy number to the base ploid. Chromosome XII demonstrates a change in copy number that is evident as a change in base ploidy and allele balance.

figures previously reported (**Figure 4**; Zhu et al., 2016) that indicated intragenomic variation in copy number. This copy number variation was demonstrated to be minor relative to the entire genome (**Figure 5**), indicating that while sample YJM1098 may be predominantly diploid, it still contains variation that would not be apparent from whole genome summaries. The use of the *vcfR* functions freq_peak() and peak_to_ploid() provided a sliding window analysis that revealed intragenomic variation in copy number. **Figure 6** demonstrated the results of the function freq_peak_plot() that revealed a sample that appeared diploid, but contains regions of low heterozygosity such that inferences cannot be made (CBS7837 chromosome XI at around 200 kbp and around 350 kbp). The sample CBS2919 appeared predominantly triploid, consistent with previous findings (Weiß et al., 2018), but also included a region on chromosome VII from its origin to around 400 kbp that appeared to have four copies. The sample CBS9564 was reported by Weiß et al. (2018) to be tetraploid, which is in agreement with our results, but also appeared to have regions on chromosome IX that had three or five copies. These findings confirm previous reports and also reveal that new information can be found by investigating specific regions within each genome.

Phytophthora infestans Dataset

The two *P. infestans* samples were sequenced at almost 200X (99189) and 300X (88069) or approximately 100X per expected chromosome (**Figure 7**; Weiß et al., 2018). The genomes had heterozygosities of around 0.003–0.006 heterozygous positions per site (**Figure 8**). Because the variant caller tends to aggressively call variants, this estimate may include false positives and therefore may be an overestimate of the true biological value.

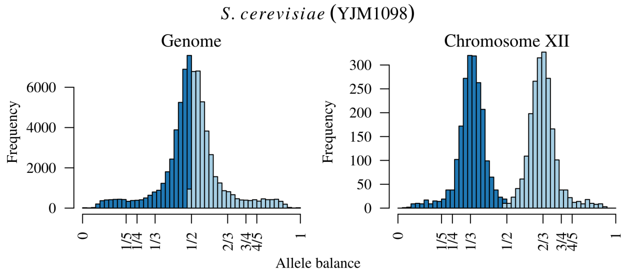


FIGURE 5 | The distribution of allele balance values for an entire sample of *Saccharomyces cerevisiae* and the distribution for just chromosome XII. Note the *y*-axis for each plot. The distribution on the right is contained within the distribution of the entire sample on the left so that this variation in copy number is hidden in plain sight.

Examination of the genomic distribution of allele balance values confirmed the report of Weiß et al. (2018) that isolate 99189 was predominantly diploid while 88069 was predominantly triploid (**Figure 9**). However, through windowing across the supercontig, we were able to observe that while isolate 99189 does appear to be predominantly diploid, a large portion of its supercontig_1.29 appears to have three copies (**Figure 10**) demonstrating previously uncharacterized intragenomic variation in copy number.

Pinfsc50 Dataset

The pinfsc50 dataset provides an opportunity to evaluate data with more moderate and more typical lower read depths. This data represents samples for a population of *P. infestans* at supercontig 50 that were sequenced between ca. 10X to 70X

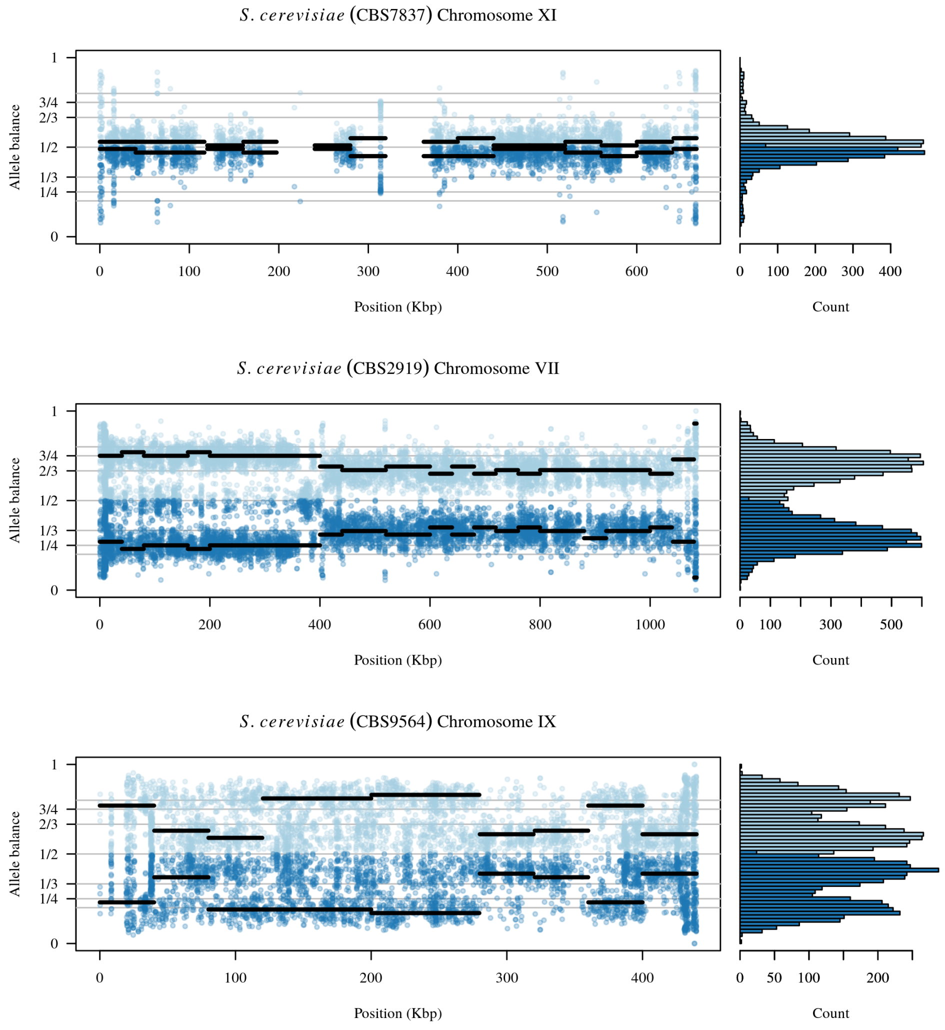


FIGURE 6 | The chromosomal distribution of heterozygous positions and their allele balance. Each plot represents one chromosome. At each variable position along the chromosome there is a pair of dots: a light blue dot above 1/2 and a darker blue dot below 1/2. These dots are the allele balance for each variant. Horizontal lines represent windows where the width is the user specified window size and the elevation is the summarized allele balance for the window. The marginal histogram summarizes the entire chromosome. The top plot is chromosome XI from sample CBS7837 and represents a diploid example. Regions at 230 and 350 kbp are regions that exhibit low levels of heterozygosity and the lack of a horizontal line indicates that these regions were omitted from the results. The middle panel is from chromosome VII of sample CBS2919. This chromosome appears to consist of four copies from its origin to around 400 kbp where it changes to three copies. The bottom panel is chromosome IX from sample CBS9564. This chromosome appears to consist of regions that have three copies as well as regions with five copies.

coverage (**Figure 11**). The distribution of allele balance values for these samples (**Figure 12**) demonstrated a range of copy numbers from diploid (e.g., strain P17777us22) to triploid (strain P13626). However, several samples (e.g., strains P1362 or t30-4) appeared to be ambiguous as to their copy number. This demonstrates that not all samples that have been sequenced from typical sequencing projects may be of suitable quality for copy number determination.

Population Differentiation

The function genetic_diff() calculates genetic differentiation for mixed copy number populations (**Table 2**). It retains the chromosome and position information from the VCF data to maintain the coordinate system. Heterozygosities as well as the number of alleles observed in each population are returned. If the number of alleles in data are unknown, this latter information may be used to summarize this information. For larger data sets, quantiles can be calculated to identify loci of unusual allele counts. The function reports G_{ST}, maximum heterozygosity, maximum G_{ST} and uses these to calculate G'_{ST}. The returned data structure is a simple data.frame which should easily facilitate further analysis and presentation of this information with the universe of R functionality.

Performance

Regression analysis revealed that execution time scaled linearly with genome size (**Figure 13**). There was a highly significant

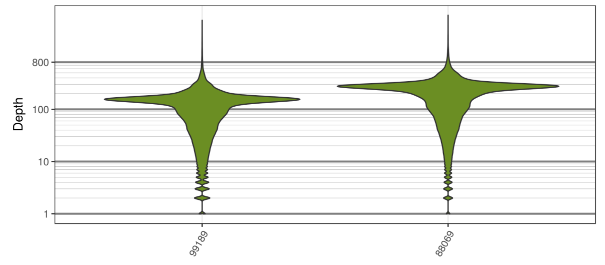


FIGURE 7 | The distribution of sequence depths at variable positions for *P. infestans* samples produced by Weiß et al. (2018). These plots are similar to the *S. cerevisiae* plots in that most of the genome appears to have been sequenced at a base ploidy level, but long tails indicate that regions above and below this level exist.

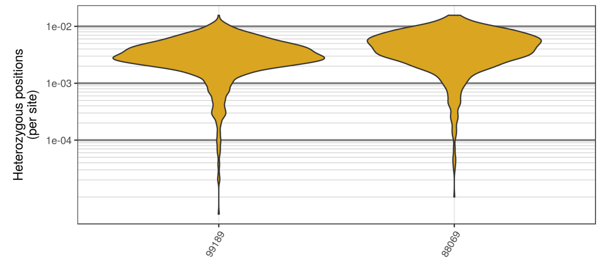


FIGURE 8 | Genomic distribution of heterozygosity among genomic windows for the two *P. infestans* samples sequenced by Weiß et al. (2018). Each genome was divided into 200 kbp windows, the number of heterozygous positions were counted, and this count was divided by the window size. The *P. infestans* genome consists of 4,921 supercontigs, many of which were below the size of these windows. In order to mitigate this, only supercontigs that resulted in at least two windows are summarized here. Note that these are raw variants from the VCF file produced by the variant caller (in our case GATK HaplotypeCaller). Because most variant callers take an aggressive perspective on variant calling, the values presented are likely an over-estimate of heterozygosity.

relationship between execution time and genome size (**Table 3**) indicating that our benchmarking may be a good predictor of how the method will perform with other genomes.

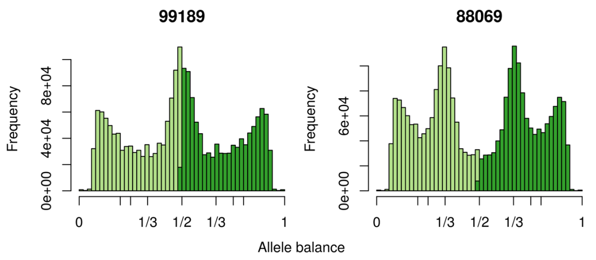


FIGURE 9 | The distribution of allele balance frequencies for samples sequenced by Weiß et al. (2018). This graphically validates the ploidy levels reported by Weiß et al. (2018).

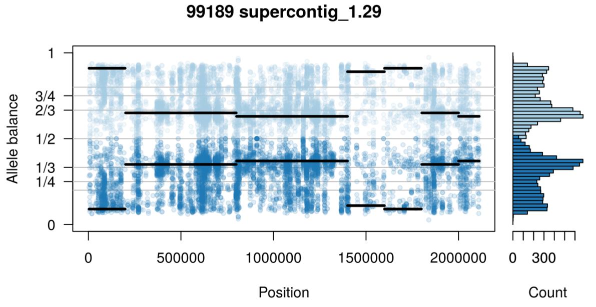


FIGURE 10 | Supercontig_1.29 of *P. infestans* isolate 99189 appears predominantly triploid in contrast to the rest of its genome that appeared to be diploid (compare with **Figure 9**). Values of 0 (no read support for the allele) and 1 (all reads support one allele) are expected to be homozygous calls. Because this is an analysis of heterozygous positions these have been omitted from this plot.

AVAILABILITY

Version 1.7.0 of the package *vcfR* had been released at the time of submission of this manuscript and contains all of the novel features described here. This version is available on CRAN (https://CRAN.R-project.org/package=vcfR) and at the Grünwald lab's GitHub site (https://github.com/grunwaldlab/vcfR). More information and example code can be found at:

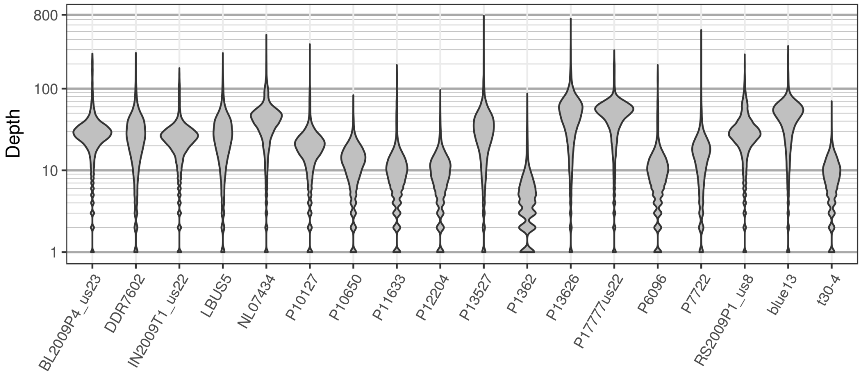


FIGURE 11 | The distribution of sequence depths at variable positions for *P. infestans* samples from the pinfsc50 dataset with variants called for supercontig 50.

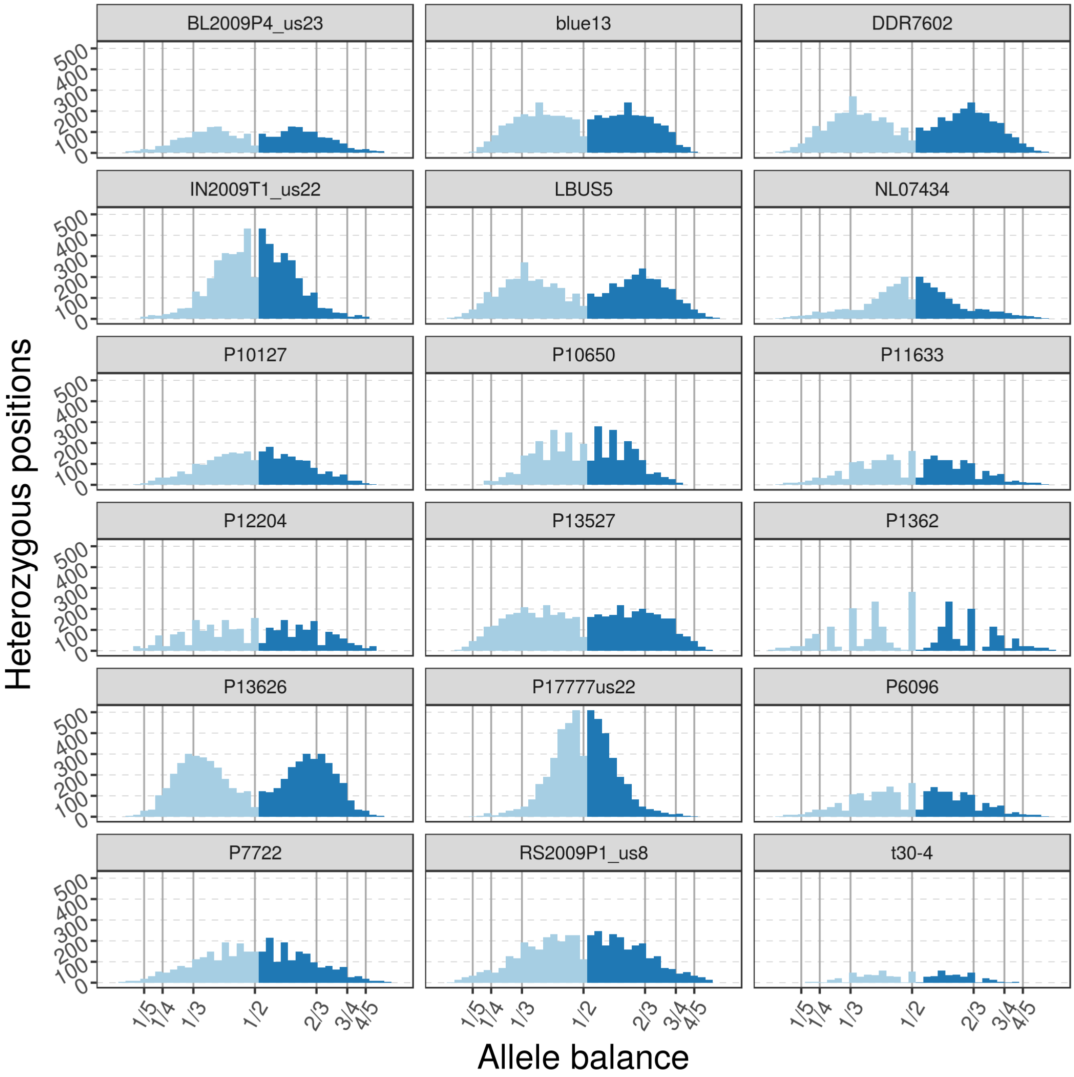


FIGURE 12 | The distribution of allele balance values for variants from supercontig_1.50 of *P. infestans*. These samples are of a more typical read depth than the other samples presented here. Note that some samples may not have a copy number that is easily determined. This illustrates the importance of providing numerical summaries as well as visualizations for the data that demonstrate edge cases as well as methods to address poor quality (e.g., removal of data based on read depth thresholds).

https://knausb.github.io/vcfR_documentation/. Data and scripts used to produce figures in this manuscript are available at the project's Open Science Framework site (Knaus and Grünwald, 2018).

REQUIREMENTS

- R version 3.0.1 or greater and *vcfR* 1.7.0.

INSTALLATION

At the R console, *vcfR* can be installed from CRAN as follows:

```
install.packages('vcfR')
library('vcfR')
```

DISCUSSION

Numerous studies have used high throughput sequencing to study genetic diversity in populations based on genotypes, or single nucleotide polymorphisms, inferred by variant callers. To our knowledge there is currently no variant caller that can infer the number of alleles to call. Instead, the investigator must specify the number of alleles to call *a priori*. Here we present novel methodologies to infer genomic and subgenomic copy number using HTS data as well as to visualize these data in the R environment.

Our method builds on existing methods by using a sliding window approach to infer copy number based on the frequency that the most abundant and second most abundant alleles were sequenced at. While we designed this method to work with VCF

TABLE 2 | Genetic differentiation as reported by the function genetic_diff().

CHROM	POS	Hs_a	Hs_b	Ht	n_a	n_b	Gst	Htmax	Gstmax	Gprimest
Supercontig_1.50	2	0.42	0.42	0.4650	20	20	0.096	0.710	0.408	0.237
Supercontig_1.50	246	0.42	0.42	0.4632	20	30	0.093	0.698	0.399	0.234
Supercontig_1.50	549	0.42	0.42	0.4600	20	40	0.0870	0.678	0.380	0.229

The chromosome (CHROM) and position (POS) are retained from the VCF data. Heterozygosities for each population (a and b) and total heterozygosity are reported. The number of alleles (n_a, n_b)_observed in each population are reported. Lastly, G_{ST}, maximum heterozygosity (Htmax), maximum G_{ST} (Gstmax) and G'_{ST} (Gprimest) are calculated.

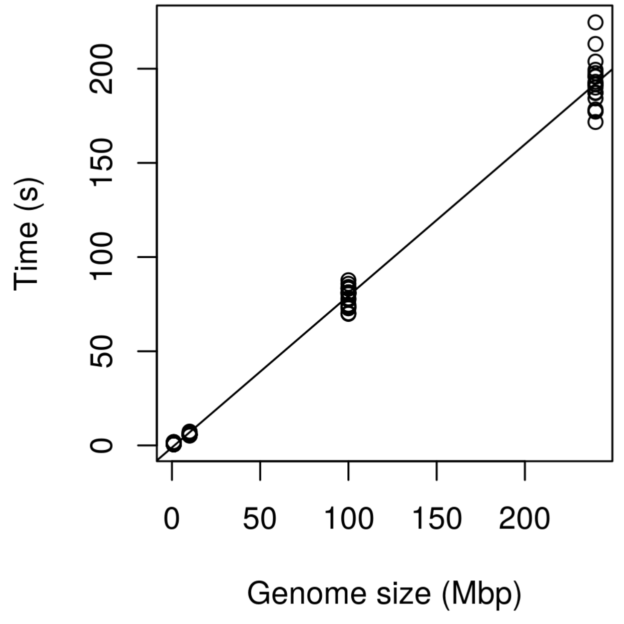


FIGURE 13 | Performance of the method expressed as execution time (seconds) as a function of genome size (Mbp). The Genome of *P. infestans* 99189 was used and subsampled at 100, 10, and 1 Mbp. Performance appears to scale linearly with the 240 Mbp genome being processed in just over 3 min.

data (Danecek et al., 2011) using the R package *vcfR* (Knaus and Grünwald, 2017), we feel an important role of our method is to help make this data available to the existing universe of R packages. VCF data only includes information on variable positions within the genome. We therefore produce a lookup table to identify which genomic windows variants belong to. Other functions convert the VCF data into numeric matrices. In theory, this information could be used to implement other functionality, such as applying mixture models (Leisch and Gruen, 2012; Fraley et al., 2012) to the data. It also means that other visualization tools available to the R environment can be used beyond those provided here. Because characterization of copy number may be challenging in certain regions of the genome, e.g., regions rich in transposable elements or problematic assemblies, we provided the count of heterozygous positions for each window as well as the distance from expectation. These metrics provide tools to help judge whether certain regions may have well predicted copy numbers or which regions may require further investigation.

TABLE 3 | Coefficients resulting from the linear regression of execution time (seconds) as a function of genome size (Mbp).

Coefficient	Estimate	Standard error	*t*-value	*P*-value
Intercept	−1.085	1.010	−1.075	0.286
Slope	0.805	0.008	103.663	<2e-16

The intercept was not significantly different from zero while the slope was highly significantly different from zero.

The existing methods most similar to ours include those of Zohren et al. (2016), Gompert and Mock (2017), and Weiß et al. (2018) because they are all based on the frequency that alleles were sequenced at. Zohren and colleagues used allele balance (which they referred to as allelic ratio) and fit beta-binomial distributions to model diploid individuals and beta-binomial mixture models (the fitting of multiple distributions to a population of data) to model triploid and tetraploid individuals. Likelihoods for each ploidy model were compared using AIC (Akaike, 1974), resulting in a single ploidy call for each sample. R code to implement their method is available at Dryad. Gompert and Mock model the ratio of the abundance of the non-reference allele (from biallelic SNPs) to the total number of reads sequenced at each variant using binomial distributions in a Bayesian framework resulting in a single ploidy call for each sample. Their method is implemented in R using *rjags* (Plummer, 2016) and is available on CRAN as the package *gbs2ploidy*. The method of Weiß and colleagues is similar to that of Zohren and colleagues in that it employs mixture models; however, it differs in that it uses Gaussian components. It also differs in that it is written in C and designed to work on the BAM files as opposed to heterozygous positions determined by a variant caller. Because it is implemented in a compiled language it is very fast relative to the R implementations. It is also unique in that it employs a uniform noise component. The sample CBS7837 in **Figure 1** has a well-defined peak, yet the base of the peak varies almost from zero to one indicating a substantial amount of data that deviates from any of our expectations. Similarly, the sample CBS2919 in **Figure 1** has two well defined peaks but the data does not go to zero between these peaks. This phenomenon can be seen in Zohren and colleagues' **Figure 2** and Yoshida et al. (2013) **Figure 8** and is part of our justification for the use of a non-parametric method. Weiß and colleagues fit this uniform component in an attempt to capture the noise in the data leaving the putatively cleaner data for their Gaussian mixture model. Their software is available on GitHub in the repository named nQuire.

The method presented has been designed to work with VCF data (Danecek et al., 2011) that contains the number of times each allele was sequenced for each variant. In theory, any method that produces a valid VCF file, or the counts of times the most abundant and second most abundant allele were sequenced in a format that can be read into R, can be analyzed. While the examples presented here are based on whole genome sequencing our method should be applicable to data generated with reduced representation libraries. For example, we've also used the method with genotyping-by-sequencing data (Elshire et al., 2011) processed with TASSEL (Bradbury et al., 2007). However, there are some practical matters to consider. This is an analysis of heterozygous positions. Homozygous positions will appear similar regardless of copy number and are uninformative. Organisms that are inbred or have a mode of reproduction that includes selfing may have a low density of heterozygous positions making inferences using our method challenging. The use of reduced representation libraries may also contribute to a lower number of observed heterozygous positions requiring use of larger windows ultimately resulting in a lower resolution to the inference of copy number variation.

There is currently a diversity of methods available for the analysis of high-throughput sequencing that demonstrates a diversity of performance. This diversity in performance exists in *de novo* assembly software (Earl et al., 2011; Bradnam et al., 2013), variant callers (Pabinger et al., 2014), copy number variation callers (Duan et al., 2013; Pabinger et al., 2014), and metagenomic pipelines (Edgar, 2017). This diversity is likely due to the nascent nature of the data and methods used to analyze it. We hope our method will contribute to the analysis of CNV, but also hope it will stimulate the development of new tools or the integration of these existing methods into new tools to explore copy number variation. Perhaps future improvements can be found by integrating sequence coverage and allele balance data as some authors have already done graphically (Zhu et al., 2016).

AUTHOR CONTRIBUTIONS

BK conceived the project, wrote code, wrote the documentation, and wrote the manuscript. NG conceived the project, coordinated the collaborative effort, discussed interpretation, wrote the manuscript, and obtained funding.

FUNDING

This research is supported in part by U.S. Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D and USDA National Institute of Food and Agriculture Grant 2011-68004-30154.

ACKNOWLEDGMENTS

Mention of trade names or commercial products in this manuscript are solely for the purpose of providing specific information and do not imply recommendation or endorsement.

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

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

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2:10. doi: 10.1186/2047-217X-2-10

Cotter, D., O'Donovan, V., O'Maoiléidigh, N., Rogan, G., Roche, N., and Wilkins, N. P. (2000). An Evaluation of the use of triploid Atlantic salmon (*Salmo salar* L.) in minimising the impact of escaped farmed salmon on wild populations. *Aquaculture* 186, 61–75. doi: 10.1016/S0044-8486(99)00367-1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Duan, J., Zhang, J.-G., Deng, H.-W., and Wang, Y.-P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 8:e59128. doi: 10.1371/journal.pone.0059128

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* 21, 2224–2241. doi: 10.1101/gr.126599.111

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. *PeerJ* 5:e3889. doi: 10.7717/peerj.3889

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:e19379. doi: 10.1371/journal.pone.0019379

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., et al. (2014). The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* 4, 389–398. doi: 10.1534/g3.113.008995

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Seattle, WA: University of Washington.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907

Gompert, Z., and Mock, K. E. (2017). Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Mol. Ecol. Resour.* 17, 1156–1167. doi: 10.1111/1755-0998.12657

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., et al. (2009). Genome sequence and analysis of the Irish Potato Famine pathogen *Phytophthora infestans*. *Nature* 461, 393–398. doi: 10.1038/nature08358

Hassold, T., and Hunt, P. (2001). To Err (Meiotically) is human: the genesis of human aneuploidy. *Nat. Rev. Genet.* 2, 280–291. doi: 10.1038/35066065

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution* 59, 1633–1638. doi: 10.1111/j.0014-3820.2005.tb01814.x

Johnstone, R. (1992). *Production and Performance of Triploid Atlantic Salmon in Scotland. Scottish Aquaculture Research Report*. Aberdeen: Marine Laboratory.

Jombart, T. (2008). Adegenet: a R Package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Katju, V., and Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front. Genet.* 4:273. doi: 10.3389/fgene.2013.00273

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40:e69. doi: 10.1093/nar/gks003

Knaus, B. J., and Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* 17, 44–53. doi: 10.1111/1755-0998.12549

Knaus, B. J., and Grünwald, N. J. (2018). *Methods for Calling Ploidy or Copy Number Variation in R*. Charlottesville, VA: Center for Open Science. doi: 10.17605/OSF.IO/ZQ879

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34, 591–602. doi: 10.1002/gepi.20516

Leisch, F., and Gruen, B. (2012). *Flexmix: Flexible Mixture Modeling. R Package Version*. Available at: https://CRAN.R-project.org/package=flexmix

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997

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

Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., et al. (2012). CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28, 1307–1313. doi: 10.1093/bioinformatics/bts146

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

McKinney, G. J., Waples, R. K., Seeb, L. W., and Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol. Ecol. Resour.* 17, 656–669. doi: 10.1111/1755-0998.12613

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York City, NY: Columbia University Press.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., et al. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* 15, 256–278. doi: 10.1093/bib/bbs086

Paradis, E., Gosselin, T., Goudet, J., Jombart, T., and Schliep, K. (2017a). Linking genomics and population genetics with R. *Mol. Ecol. Resour.* 17, 54–66. doi: 10.1111/1755-0998.12577

Paradis, E., Gosselin, T., Grünwald, N. J., Jombart, T., Manel, S., and Lapp, H. (2017b). Towards an integrated ecosystem of R packages for the analysis of population genetic data. *Mol. Ecol. Resour.* 17, 1–4. doi: 10.1111/1755-0998.12636

Plummer, M. (2016). *Rjags: Bayesian Graphical Models Using MCMC*. Available at: https://CRAN.R-project.org/package=rjags

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., et al. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature* 530, 177–183. doi: 10.1038/nature16549

Todd, R. T., Forche, A., and Selmecki, A. (2017). Ploidy variation in fungi: polyploidy, aneuploidy, and genome evolution. *Microbiol. Spectr.* 5:FUNK-0051-2016.

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

Varoquaux, F., Blanvillain, R., Delseny, M., and Gallois, P. (2000). Less is better: new approaches for seedless fruit production. *Trends Biotechnol.* 18, 233–242. doi: 10.1016/S0167-7799(00)01448-7

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., and Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19:122. doi: 10.1186/s12859-018-2128-z

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., et al. (2013). The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish Potato Famine. *Elife* 2:e00731. doi: 10.7554/eLife.00731

Zhu, Y. O., Sherlock, G., and Petrov, D. A. (2016). Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)* 6, 2421–2434. doi: 10.1534/g3.116.029397

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., et al. (2016). Unidirectional diploid–tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Mol. Ecol.* 25, 2413–2426. doi: 10.1111/mec.13644

Conflict of Interest Statement: 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 Plant Science

REVIEW
published: 18 April 2018
doi: 10.3389/fpls.2018.00513

Tools for Genetic Studies in Experimental Populations of Polyploids

*Peter M. Bourke, Roeland E. Voorrips, Richard G. F. Visser and Chris Maliepaard**

Plant Breeding, Wageningen University & Research, Wageningen, Netherlands

Polyploid organisms carry more than two copies of each chromosome, a condition rarely tolerated in animals but which occurs relatively frequently in the plant kingdom. One of the principal challenges faced by polyploid organisms is to evolve stable meiotic mechanisms to faithfully transmit genetic information to the next generation upon which the study of inheritance is based. In this review we look at the tools available to the research community to better understand polyploid inheritance, many of which have only recently been developed. Most of these tools are intended for experimental populations (rather than natural populations), facilitating genomics-assisted crop improvement and plant breeding. This is hardly surprising given that a large proportion of domesticated plant species are polyploid. We focus on three main areas: (1) polyploid genotyping; (2) genetic and physical mapping; and (3) quantitative trait analysis and genomic selection. We also briefly review some miscellaneous topics such as the mode of inheritance and the availability of polyploid simulation software. The current polyploid analytic toolbox includes software for assigning marker genotypes (and in particular, estimating the dosage of marker alleles in the heterozygous condition), establishing chromosome-scale linkage phase among marker alleles, constructing (short-range) haplotypes, generating linkage maps, performing genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses, and simulating polyploid populations. These tools can also help elucidate the mode of inheritance (disomic, polysomic or a mixture of both as in segmental allopolyploids) or reveal whether double reduction and multivalent chromosomal pairing occur. An increasing number of polyploids (or associated diploids) are being sequenced, leading to publicly available reference genome assemblies. Much work remains in order to keep pace with developments in genomic technologies. However, such technologies also offer the promise of understanding polyploid genomes at a level which hitherto has remained elusive.

Keywords: polyploid genetics, polyploid software tools, autopolyploid, allopolyploid, segmental allopolyploid

OPEN ACCESS

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Keiichi Mochida,
RIKEN Center for Sustainable Resource Science (CSRS), Japan
Patricio Ricardo Munoz,
University of Florida, United States

****Correspondence:***
Chris Maliepaard
chris.maliepaard@wur.nl

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

Received: *25 January 2018*
Accepted: *04 April 2018*
Published: *18 April 2018*

Citation:
Bourke PM, Voorrips RE, Visser RGF and Maliepaard C (2018) Tools for Genetic Studies in Experimental Populations of Polyploids. Front. Plant Sci. 9:513. doi: 10.3389/fpls.2018.00513

INTRODUCTION

One of the most fundamental descriptions of any organism is its ploidy level and chromosome number, generally written in the form $2n = 2x = 10$ (here, for the ubiquitous model plant species *Arabidopsis thaliana* L.). Plant scientists in particular will be familiar with this representation of the chromosomal constitution of the sporophyte generation (i.e., the adult plant). The second term in this seemingly simple equation describes the normal complement of chromosomal copies possessed by a member of that species, which is generally 2× ("two times") for diploids. Species where this number exceeds two are collectively referred to as polyploids. Not unexpectedly, each polyploid individual is the product of the fusion of gametes from two parents, just like their diploid counterparts. In other words, polyploids can also be defined as individuals derived from non-haploid gametes (in the case of triploids derived from diploid × tetraploid crosses, only one gamete satisfies this condition). The transmission of non-haploid gametes is one of the main "complexifying" features of polyploidy, leading to a whole range of implications for the genetic analysis of these "hopeful monsters" (Goldschmidt, 1933).

The ongoing genomics revolution can be seen as a rising tide which has also lifted the polyploid genetics boat, although not quite to the same level as for diploids. Most genetic advances are made in model organisms, among which self-fertilizing diploid species predominate. It is therefore not surprising that most tools and techniques for molecular-genetic studies are specific to diploids. However, polyploid species are particularly important to mankind in the provision of food, fuel, feed, and fiber (not to mention "flowers," if ornamental plant species are also included), making the genetic analysis of polyploid species an important avenue of research for crop improvement.

Although a collective term such as "polyploidy" has its uses, it tends to obscure some fundamental differences between its members. For example, polyploids are generally subdivided into autopolyploids and allopolyploids (Kihara and Ono, 1926). Autopolyploids arise through genomic duplication within a single species, generally through the production of unreduced gametes (Harlan and De Wet, 1975) and exhibit polysomic inheritance, meaning pairing and recombination can occur between all homologous copies of each chromosome during meiosis. One of the most well-studied examples is autotetraploid potato (*Solanum tuberosum* L.). Allopolyploids, on the other hand, are the product of genomic duplication between species [usually through hybridisation involving unreduced gametes (Harlan and De Wet, 1975)] and display disomic inheritance, where more-related chromosome copies ("homologs") may pair and recombine during meiosis, whilst less-related chromosome copies ["homoeologs," also spelled "homeologs" (Glover et al., 2016)] do not. Among allopolyploids, allohexaploid wheat (*Triticum aestivum* L.) is probably the most well-studied. If pairing and recombination between homoeologs occurs to a limited extent, the species may be referred to as "segmental allopolyploid" (Stebbins, 1947), traditionally deemed to have arisen from hybridisation between very closely related species (Stebbins, 1947; Chester et al., 2012) but which may also be the result of partially diploidised autopolyploidy (Soltis et al., 2016). In many cases, a species cannot be clearly designated as one type or another, leading to uncertainty or debate on the subject (Barker et al., 2016; Doyle and Sherman-Broyles, 2016). From the perspective of genetics and inheritance, allopolyploids behave much like diploid species and therefore many of the tools developed for diploids can be directly applied. The main challenge that faces allopolyploid geneticists is in distinguishing between homoeologous gene copies carried by sub-genomes within an individual (Kaur et al., 2012; van Dijk et al., 2012; Rothfels et al., 2017). Autopolyploids (and segmental allopolyploids) do not behave like diploids, and are therefore in most need of specialized methods and tools for subsequent genetic studies. In this review we focus primarily on the availability of tools and resources amenable to polysomic [and "mixosomic" (Soltis et al., 2016)] species, with less emphasis on allopolyploid-specific solutions. Although the development of novel methodologies for the genetic analysis of polyploids are interesting, without translation into a software tool for use by the research community they remain purely conceptual and with limited impact. We therefore try to limit our attention to the tools currently available rather than cataloging descriptions of unimplemented methods.

Experimental populations, in use since Mendel's ground-breaking work (Mendel, 1866), are traditionally derived from a controlled cross between two parental lines of interest (either directly studying the F_1 or some later generation). We use the term here to distinguish our subject matter from "wild" or "natural" populations, which would necessitate sampling individuals from an extant population in the wild. Quantitative genetics, particularly the genetics of human pathology, has greatly benefitted from the use of large panels of individuals to perform so-called "genome-wide association studies" (GWAS). The use of such panels offers to complement the experimental toolbox of polyploid geneticists as well, and although perhaps not strictly speaking an "experimental" population, we consider them relevant to the current discussion.

Here, we review three main areas: (1) polyploid genotyping, including the scoring of marker dosage (allele counts) and generation of haplotypes; (2) genetic and physical mapping, where we look at the possibilities for linkage mapping as well as the availability of reference sequences; and (3) quantitative trait analysis and genomic selection, including tools that perform quantitative trait locus (QTL) analysis in bi-parental populations, genome-wide association analysis (GWAS) and genomic selection and prediction. We also consider the current tools to simulate polyploid organisms for *in silico* studies, as well as those that can help determine the mode of inheritance of the species being studied. We reflect on current and future developments, and the tools that will be needed to keep pace with the innovations we are witnessing in genomic technologies.

POLYPLOID GENOTYPING

One of the most crucial aspects in the study of polyploid genetics is the generation of accurate genotypic data. However, it is also fraught with difficulties, not least the detection of multiple loci when only a single locus is targeted (Mason, 2015; Limborg et al., 2016). Various technologies exist, with almost all current applications aimed at identifying single nucleotide polymorphisms (SNPs). Although many genomic "service-providers" (e.g., companies or institutes that offer DNA sequencing) have their own tools to analyze and interpret raw data, these tools are not always suitable for use with polyploid datasets. Gel-based marker technologies continue to be used and retain certain advantages (e.g., low costs associated with small marker numbers, requiring only basic laboratory facilities, multi-allelism etc.). However, most studies now rely on SNP markers for genotyping due to their great abundance over the genome, their high-throughput capacity and their low cost per data point. Targeted genotyping such as SNP arrays (a.k.a. "SNP chips") rely on previously identified and selected polymorphisms, usually identified from a panel of individuals chosen to represent the gene pool under investigation. In contrast, untargeted genotyping generally uses direct sequencing of individuals, albeit after some procedure to reduce the amount of DNA to be sequenced [e.g., by exome sequencing (Ng et al., 2009) or target enrichment (Mamanova et al., 2010)]. The disadvantages of targeted approaches have been well explored (particularly regarding ascertainment bias, where the set of targeted SNPs on an array poorly represents the diversity in the samples under investigation due to biased methods of SNP discovery) (Albrechtsen et al., 2010; Moragues et al., 2010; Didion et al., 2012; Lachance and Tishkoff, 2013), although there are advantages and disadvantages to both methods (Mason et al., 2017). Apart from costs, differences exist in the ease of data analysis following genotyping, with sequencing data requiring greater curation and bioinformatics skills (Spindel et al., 2013; Bajgain et al., 2016) as well as potentially containing more erroneous and missing data (Spindel et al., 2013; Jones et al., 2017).

In polyploids, SNP arrays have been developed in numerous species [recently reviewed by (You et al., 2018)], which include both autopolyploid (or predominantly polysomic polyploids) and allopolyploid species. Examples of the former include alfalfa (Li et al., 2014), chrysanthemum (van Geest et al., 2017b), potato (Hamilton et al., 2011; Felcher et al., 2012; Vos et al., 2015), rose (Koning-Boucoiran et al., 2015) and sour cherry (Peace et al., 2012). Examples of allopolyploid SNP arrays include cotton (Hulse-Kemp et al., 2015), oat (Tinker et al., 2014), oilseed rape (Dalton-Morgan et al., 2014; Clarke et al., 2016), peanut (Pandey et al., 2017), strawberry (Bassil et al., 2015) and wheat (Akhunov et al., 2009; Cavanagh et al., 2013; Wang et al., 2014; Winfield et al., 2016). Untargeted approaches such as genotyping using next-generation sequencing have also been applied, for example in autopolyploids such as alfalfa (Zhang et al., 2015; Yu et al., 2017), blueberry (McCallum et al., 2016), bluestem prairie grass (*Andropogon gerardii*) (McAllister and Miller, 2016), cocksfoot (*Dactylis glomerata*) (Bushman et al., 2016), potato (Uitdewilligen et al., 2013; Sverrisdóttir et al., 2017), sugarcane (Balsalobre et al., 2017; Yang et al., 2017b) and sweet potato (Shirasawa et al., 2017), and in allopolyploids such as coffee (Moncada et al., 2016), cotton (Islam et al., 2015; Reddy et al., 2017), intermediate wheatgrass (*Thinopyrum intermedium*) (Kantarski et al., 2017), oat (Chaffin et al., 2016), prairie cordgrass (*Spartina pectinata*) (Crawford et al., 2016), shepherd's purse (*Capsella bursa-pastoris*) (Cornille et al., 2016), wheat (Poland et al., 2012; Edae et al., 2015), and zoysiagrass (*Zoysia japonica*) (McCamy et al., 2018) (noting that the precise classification of some of these species as auto- or allopolyploids has yet to be conclusively determined). Whatever the technology used, it is clear that we are currently witnessing an explosion of interest in polyploid genomics. However, the critical issue of how to make sense of this data remains, starting with the assignment of marker dosage, a.k.a. "genotype calling."

Assignment of Dosage

One of the key distinguishing features of polysomic polyploidy is the fact that there are multiple heterozygous conditions possible in genotyping data. We use the term marker "dosage" to denote the minor allele count of a marker; a species of ploidy q possesses $q + 1$ distinct dosage classes in the range 0 to q (**Figure 1**). Of course the concept of marker dosage could also be used in diploid species, but coding systems such as the lm × ll / nn × np / hk × hk system (Van Ooijen, 2006) predominate. Marker dosage is generally understood to apply to bi-allelic markers (such as single SNPs), although it is conceivable to score marker dosage at multi-allelic loci. If marker dosage cannot be accurately assessed, genotypes would likely have to be dominantly scored (i.e., all heterozygous classes would be grouped with one of the homozygous classes), resulting in a loss of information (Piepho and Koch, 2000).

All available dosage-calling tools rely on a population in order to determine marker dosage. In other words, calibration between the various dosage classes is performed across the population (for which we are not implying any degree of relatedness in the population other than coming from the same species). All current tools are designed to process genotyping data from SNP arrays, using the relative strength of two allele-specific (fluorescent) signals to assign a discrete dosage value. With increasing interest in genotyping using next generation sequencing (GNGS), we anticipate that tools which use read-counts of potentially multiple SNPs (or multi-SNP haplotypes) will soon be developed, although these have yet to appear. One of the current challenges under investigation regarding GNGS-based genotype calling is the accurate determination of dosage (Kim et al., 2016), which may require relatively deep sequencing [e.g., 60–80 × coverage estimated in autotetraploid potato (Uitdewilligen et al., 2013)].

Returning to the SNP array-based tools, the two main service providers for high-density SNP arrays, Illumina and Affymetrix, both offer proprietary software solutions for analyzing polyploid datasets. Affymetrix's Power Tools and Illumina's GenomeStudio (with its Polyploid Genotyping Module) have both been developed with both diploid and polyploid datasets in mind. However, there have also been a number of genotyping tools

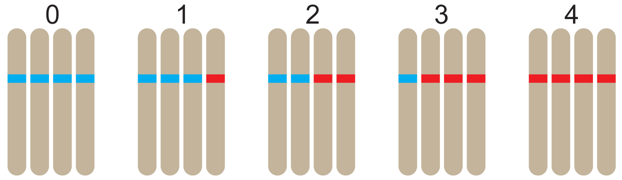


FIGURE 1 | In a tetraploid, five distinct dosages are possible at a bi-allelic marker positions, ranging from 0 copies of the alternative allele through to 4 copies. Here, the alternative allele is colored red, with the reference allele colored blue.

that have been put into the public domain. One of the first of these to be released was fitTetra (Voorrips et al., 2011), a freely available R package (R Core Team, 2016) designed to assign genotypes to autotetraploids that were genotyped on either Illumina's Infinium or Affymetrix's Axiom arrays. fitTetra fits mixture models to bi-allelic SNP intensity ratios either under the constraint of Hardy-Weinberg equilibrium within the population, or as an unconstrained fit, using an expectation-maximization (EM) algorithm in fitting. This can have the drawback of requiring significant computational resources for high-density marker datasets, although it is automated and can therefore process large datasets in a single run. The original release was specific to tetraploid data only. However, an updated version (fitPoly) can process genotyping data of all ploidy levels and has recently made available as a separate R package on CRAN[1]. The SuperMASSA application (Serang et al., 2012) can also process data from all ploidy levels (as it was initially developed to dosage-score sugarcane data, notorious for its cytogenetic complexity) and is currently hosted online by the Statistical Genetics Laboratory in the University of São Paulo, Brazil. One of the interesting features of SuperMASSA is that prior knowledge of the exact ploidy level is not needed (useful for a crop like sugarcane). Instead, the genotype configuration which maximizes the posterior probability across all specified ploidy levels is chosen. In practice, most researchers will already know the ploidy of their samples (although aneuploid progeny in some species may occur) and can constrain the model search. A drawback of the online implementation is that markers are analyzed one-by-one, and results need to be copied from the webpage each time. However, a command-line version of SuperMASSA is currently under development.

The R package polysegRatioMM (Baker et al., 2010) generates marker dosages for dominantly scored markers using the JAGS software (Plummer, 2003) for Markov Chain Monte Carlo (MCMC) generation. Fully polysomic behavior is assumed, and segregation ratios of marker data are used to derive the most likely parental scores. Although able to process data from all even ploidy levels, the software only considers a subset of marker types (marker that are nulliplex in one parent or simplex in both parents). Nowadays, there is a move away from dominantly scored markers to co-dominant marker technologies like SNPs, and parental samples are usually included in multiple replicates (and so can be genotyped directly with offspring, rather than imputed from the offspring). The package is therefore of questionable use for modern genotyping datasets. An unrelated R package, beadarrayMSV (Gidskehaug et al., 2010), was developed to handle Illumina Infinium SNP array data from "diploidising" tetraploid species such as the Atlantic salmon. The software was designed to score markers which target multiple loci (so-called multi-site variants, or MSVs), as well as single-locus markers displaying disomic inheritance. In a comparison with fitTetra, beadarrayMSV was unable to accurately genotype autotetraploid data from potato, although conversely fitTetra performed poorly on salmon data (Voorrips et al., 2011). This demonstrates that appropriate software is needed for specific situations (indeed, in many cases specific scenarios have motivated the development of specialized software).

Having prior knowledge about the expected meiotic behavior of the species is always advantageous when it comes to analyzing any polyploid data. This is especially true for the latest dosage-calling software to be released, the ClusterCall package for R (Schmitz Carley et al., 2017). Here, prior knowledge of the meiotic behavior of the species is required, since the expected segregation ratios of an F_1 autotetraploid population are used to assign dosage scores to the clusters identified through hierarchical clustering. In well-behaved autotetraploids such as potato (Swaminathan and Howard, 1953; Bourke et al., 2015) this is arguably not a problem (as long as skewed segregation does not occur), and indeed can lead to increased accuracy in genotype calling (Schmitz Carley et al., 2017). However, in less well-characterized species such as leek, alfalfa, or many ornamental species, the precise meiotic behavior may not always follow the expected tetrasomic model, causing potential problems with fitting. The authors are aware of this and suggest that alternatives like fitTetra or SuperMASSA be used in circumstances where a tetrasomic model no longer holds. Unfortunately, such prior knowledge is not always available before genotyping takes place – meiotic behavior can even differ between individuals of a species that was thought to display meiotic homogeneity (e.g., complete tetrasomy) (Bourke et al., 2017).

Haplotype Assembly

Although bi-allelic SNP markers have many practical advantages, they carry less inheritance information than multi-allelic markers. Crop researchers and breeders often wish to develop a simple diagnostic marker test for a trait of interest. Unfortunately, the chances of having a single SNP in complete linkage disequilibrium with a favorable or causative allele of a gene of interest is very small. Markers which have been found to uniquely "tag" a favorable allele in one population may not do so in another. For more than a decade, the increased power of haplotype-based associations have been known and reported in human genetic studies (Zhang et al., 2002; de Bakker et al., 2005), with the term "haplotype" denoting a unique stretch of sequence. Translating haplotyping approaches from diploid to polyploid species has been a non-trivial exercise, requiring novel algorithms to handle the overwhelming range of possibilities that can arise [especially when allowing for sequencing errors and (possible) recombinations]. Multi-SNP haplotypes can be assembled from single dosage-scored SNPs (originating from SNP array data), although haplotypes are more commonly generated using overlapping sequence reads (**Figure 2**).

A number of different polyploid haplotyping tools (for sequence reads) have been developed in recent years, including polyHap (Su et al., 2008), SATlotyper (Neigenfind et al., 2008), HapCompass (Aguiar and Istrail, 2013), HapTree (Berger et al., 2014), SDhaP (Das and Vikalo, 2015), SHEsisplus (Shen et al., 2016), and TriPoly (Motazedi et al., unpublished). Three of these tools (HapCompass, HapTree, and SDhaP) were recently compared and evaluated over a range of different simulated read depths, ploidy levels and insert sizes for paired-end reads (Motazedi et al., 2017). The authors found that each of

[1] https://cran.r-project.org/package=fitPoly

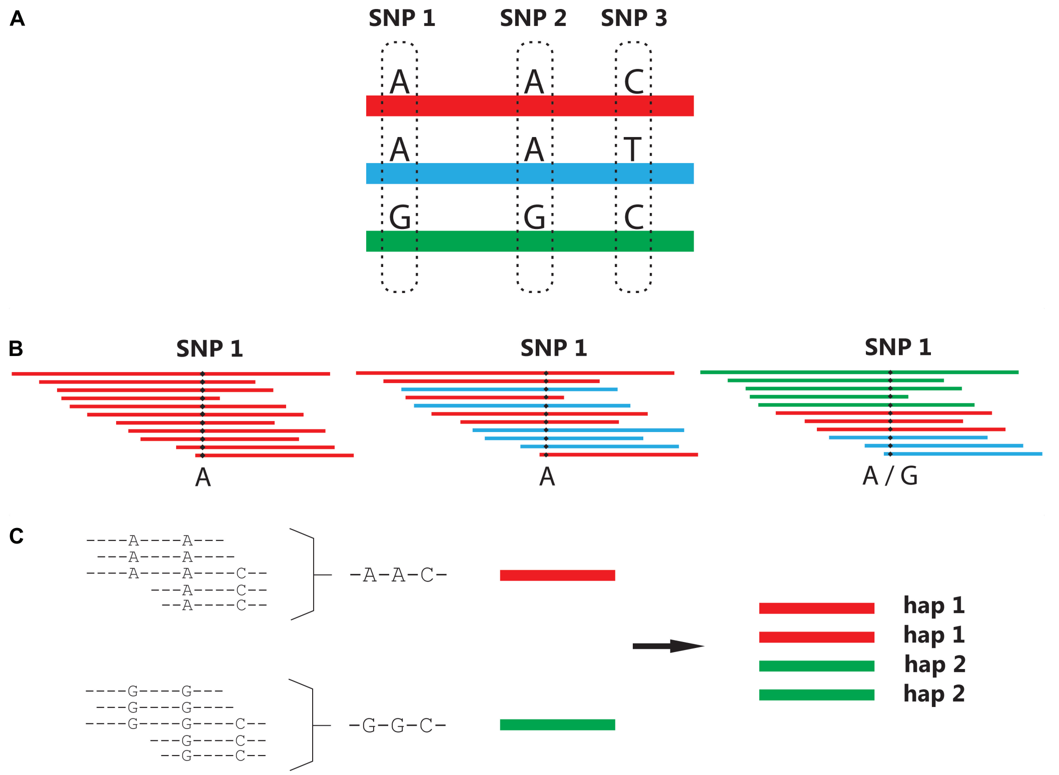


FIGURE 2 | Generation of multi-SNP haplotypes. **(A)** In this example, three possible haplotypes exist spanning polymorphic positions SNP 1, 2, and 3. **(B)** Single-SNP genotyping cannot distinguish between the "A" allele originating from different haplotypes, combining them into a single allele as illustrated in the second SNP call. **(C)** In a haplotyping approach, overlapping reads are used to re-assemble and phase single SNP genotypes. Here, the known ploidy level of the species (4×) is used to impute the dosage of the two haplotypes identified in this individual, given a 1:1 ratio between the assembled haplotype read-depths.

these software programs had particular advantages, for example HapTree was found to produce more accurate haplotypes for triploid and tetraploid data, whilst HapCompass performed best at higher ploidies (6× and higher) (Motazedi et al., 2017). Both SHEsisplus and TriPoly have yet to be independently tested. For allopolyploid species, the user-friendly Haplotag software has been designed to identify both single SNPs and multi-SNP haplotypes from genotypes developed using next generation sequencing data (Tinker et al., 2016). An interesting feature is the use of a simple "heterozygosity filter" that excludes haplotypes with higher than expected heterozygosity across a population (suggesting paralogous loci). Currently, however, data from outcrossing or autopolyploid species is not suitable for this software.

The input data of haplotyping software can be grouped into two types. Individual SNP genotyping data (with a known marker order) was used by the first wave of polyploid haplotyping implementations such as polyHap and SATlotyper. More recently, haplotyping tools use sequence reads as their input, although some pre-processing is required: reads must first be aligned followed by extraction of their SNPs (i.e., masking of non-polymorphic sites) to generate a SNP-fragment matrix with individual reads as rows and SNP positions as columns [as described for HapCompass (Aguiar and Istrail, 2013)]. In other words, all haplotyping tools [apart perhaps from Haplotag (Tinker et al., 2016)] require that users possess a certain level of bioinformatics skills. Although we expect polyploid haplotypes to become increasingly used in the future, the development of user-friendly and computationally efficient tools is first needed before haplotype-based genotypes become truly mainstream.

One interesting development is the application of haplotyping to whole genome assemblies (as opposed to genotyping a population). This has recently been attempted in the tuberous hexaploid crop sweet potato (*Ipomoea batatas*) (Yang et al., 2017a). The authors first produced a consensus assembly to which reads were re-mapped for variant calling, followed by a phasing algorithm which resolved the six haplotypes of the sequenced cultivar for about 30% of the assembly (Yang et al., 2017a). Ultimately, about half of the assembled genome could be haplotype-resolved. Future sequencing (or re-sequencing) efforts in polyploid species should produce more phased genomes, which will no doubt be useful for haplotyping applications (for example in validating predicted haplotypes).

GENETIC AND PHYSICAL MAPPING OF POLYPLOID GENOMES

One of the first steps in understanding the genetic composition of any species is the development of a map, be it a genetic map based on information about linkage and co-inheritance of specific DNA locations, or a physical map giving a reference DNA sequence for the species. In polyploid species, numerous technical and methodological complications arise that make the mapping of polyploids a much more complex endeavor than diploid mapping. However, there is currently an upsurge in interest in polyploid mapping, which has led to much progress in recent years.

Linkage Maps

Although the first genetic linkage map was developed more than 100 years ago (Sturtevant, 1913), their use in genetic and genomic studies has persisted into the "next-generation" era. This can be attributed to a number of factors. A linkage map is a description of the recombination landscape within a species, usually from a single experimental cross of interest. For breeders, knowledge of genetic distance is arguably more important than physical distance, as it reflects the recombination frequencies in inheritance studies as well as describing the extent of linkage drag around loci of interest. Many software for performing QTL analysis require linkage maps of the markers, not physical maps. This is because co-inheritance of markers and phenotypes within a population are assumed to be coupled – a physical map gives less precise information about the co-inheritance of markers than a linkage map does since physical distances do not directly translate to recombination frequencies (particularly in the pericentromeric regions). Another reason why linkage maps continue to be developed is that they are often the first genomic representation of a species, upon which more advanced representations can be built. They provide useful long-range linkage information over the whole chromosome which is often missing from assemblies of short sequence reads. This fact has been repeatedly exploited in efforts at connecting and correctly orientating scaffolds during genome assembly projects (Bartholomé et al., 2015; Fierst, 2015).

As mentioned in the Introduction, polyploids can be divided into disomic or polysomic species, with the additional possibility of a mixture of both inheritance types in the case of segmental allopolyploids. Many linkage maps in polyploids have been based exclusively on 1:1 segregating markers, also known as simplex markers [because the segregating allele is in simplex condition (one copy) in one of the parents only]. These markers possess a number of advantages over other marker segregation types, but also some distinct disadvantages. In their favor, coupling-phase simplex markers in polyploid species behave just like they would in diploid species, regardless of the mode of inheritance involved (repulsion-phase recombination frequency estimates are not invariant across ploidy levels or modes of inheritance, but exert less influence on map construction due to lower LOD scores). The advantage of this is clear: in unexplored polyploid species for which the mode of inheritance is uncertain, simplex markers allow an "assumption-free" linkage map to be created, following which the mode of inheritance can be further explored. The only exception to this is if double reduction occurs, i.e., when a segment of a single chromosome gets transmitted with its sister chromatid copy to an offspring, a consequence of multivalent pairing and a particular sequence of segregation and division during meiosis (Haldane, 1930; Mather, 1935). Double reduction occurs randomly in polysomic species and only introduces a small bias into recombination frequency estimates (Bourke et al., 2015). This means that, ignoring the possible influence of double reduction, diploid mapping software can generally be used for simplex marker sets at any ploidy level and for any type of meiotic pairing behavior (**Figure 3**), opening up a very wide range of diploid-specific software options (Cheema and Dicks, 2009).

However, simplex marker sets have some limitations. Firstly, in selecting only simplex markers, a large proportion of markers with different segregation patterns are not used. This usually reduces the map coverage (while increasing the per-marker costs of the final set of mapped markers). More importantly, simplex markers give limited information about linkage in repulsion phase, particularly at higher ploidy levels (van Geest et al., 2017a). This means that homolog-specific maps can be produced, but they are unlikely to be well-integrated between homologs in a single parent, and impossible to integrate across parents. In other words, the chromosomal numbering will most likely be inconsistent between parental maps if only simplex markers are used. Producing a consensus or fully integrated map is desirable for many reasons, including being able to detect and model more complex QTL configurations than just simplex QTL. Therefore, a truly polyploid linkage mapping tool should be able to include all marker segregation types, not just 1:1 segregating markers.

Polyploid Linkage Mapping Software

Linkage mapping can be broken into three steps – linkage analysis, marker clustering and marker ordering. There are still relatively few software tools that can perform all three of these steps for polysomic species. Perhaps the most well-known and widely used software tool is TetraploidMap for Windows (Hackett and Luo, 2003; Hackett et al., 2007). As well as producing linkage maps for autotetraploid species, this software also performs QTL interval mapping (returned to later). Recently, TetraploidMap was updated to enable the use of dosage-scored SNP data (Hackett et al., 2013). The updated version, TetraploidSNPMap (Hackett et al., 2017), is freely available to download from the Scottish BioSS website[2], and possesses a sophisticated graphical user interface (GUI) which will be extremely welcome for users in both the research and breeding community. Apart from its dependency on the Windows platform, the main drawback of TetraploidSNPMap (TSNPM) is that it is programmed to analyze autotetraploid data only, and there is no indication when or if it will be expanded to other ploidy levels or modes of inheritance. However, tetraploidy is the most common polyploid condition (Comai, 2005) and therefore this software is still relevant for a broad range of species.

Recently, an alternative linkage mapping package called polymapR was released, which is described in a pre-print manuscript (Bourke et al., unpublished). Like TSNPM, polymapR used dosage-scored marker information from F_1 populations to estimate recombination frequencies by maximum likelihood in a two-point linkage analysis. It can perform linkage analysis

[2] https://bioss.ac.uk/knowledge/tetramap.html

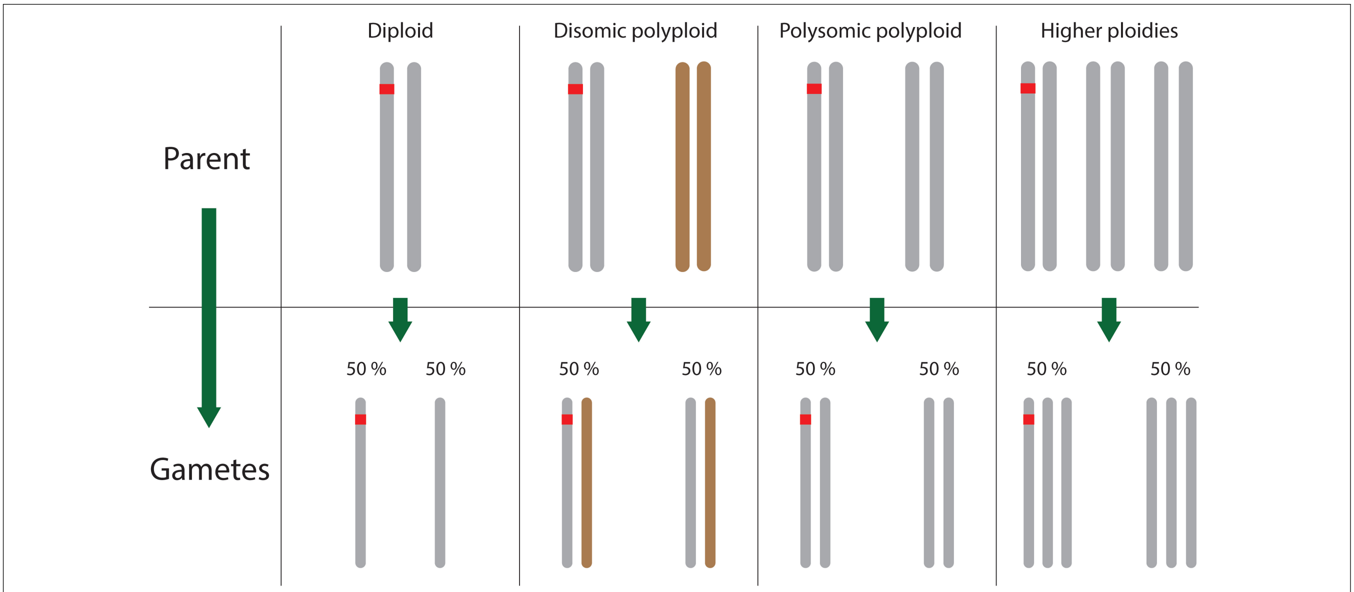


FIGURE 3 | Simplex markers (carrying a single copy of the segregating marker allele) inherit similarly across all ploidy levels and pairing behaviors, allowing diploid mapping software to be used. Here, the (simplex) SNP allele is colored red.

for polysomic triploids, tetraploids and hexaploids as well as segmental allotetraploid populations. As an R-based package it requires some level of user familiarity with R, but comes with a descriptive vignette which should make it accessible even to novice R users. It uses the same high-speed map ordering algorithm as TSNPM, namely MDSMap (Preedy and Hackett, 2016), and produces both integrated and phased linkage maps (i.e., separate maps for each parental homolog that are also integrated into a single consensus map). So far, developmental versions of this software have been used to generate high-density linkage maps in tetraploid potato (Bourke et al., 2016), tetraploid rose (Bourke et al., 2017), and hexaploid chrysanthemum (van Geest et al., 2017a).

Another recently released R package that can perform linkage map construction is the netgwas package, also described in a pre-print manuscript (Behrouzi and Wit, 2017a). netgwas claims to be able to construct maps at any ploidy level in both inbred and outbred bi-parental populations, and rather than computing recombination frequencies and LOD scores, it uses conditional dependence relationships between markers based on discrete graphical models. The algorithm automatically detects linkage groups (which are traditionally identified by a user-specified LOD threshold) and does not rely on knowledge of parental dosage scores (which should offer robustness against parental genotyping errors). The output of netgwas is clustered and ordered marker names, but without assigning genetic positions (centiMorgans) or marker phasing, which are part of the TSNPM and polymapR output. The lack of marker phasing in particular is a major drawback, as phase considerations are crucial in polyploid genetic analyses. However, given its novel and computationally efficient approach to map construction, it appears to be a very interesting addition to the current range of polyploid mapping tools.

Another software program that is able to perform all three major steps in polyploid linkage mapping is the PERGOLA package in R (Grandke et al., 2017). This software can analyze marker data from all ploidy levels and modes of inheritance, but is limited to populations derived from completely inbred (homozygous) founder parents, such as F_2 or BC_1 populations. While these sorts of experimental population are common in diploid plant species, they are much less common in polyploids due to the difficulty in reaching homozygosity through selfing (Haldane, 1930). Generally speaking, polyploids are more heterozygous than diploids (Soltis and Soltis, 2000) although there is no general consensus regarding their tolerance of inbreeding (Krebs and Hancock, 1990; Soltis and Soltis, 2000; Galloway et al., 2003; Galloway and Etterson, 2007). There are indications that polyploid plant species self-fertilize more often than their diploid relatives (Barringer, 2007). However, regardless of whether polyploids tolerate some levels of inbreeding or not, heterozygosity is maintained for many more generations in repeatedly selfed polyploids than in selfed diploids (**Figure 4**). It therefore appears likely that PERGOLA was developed for newly formed polyploids derived from inbred diploid lines. The complexities facing extant (or heterozygous) polyploid species such as unknown marker phasing, or variable marker information contents are ignored by PERGOLA, making it doubtful that this tool will have a wide impact on linkage mapping in existing polyploid populations.

One final software that should be mentioned is PolyGembler, recently described in a pre-print manuscript (Zhou et al., unpublished). It proposes a novel approach to the creation of linkage maps in outcrossing polyploids, and is also suitable for diploid mapping. Interestingly, it combines a haplotyping algorithm [derived from the polyHap algorithm (Su et al., 2008)] to first generate phased multi-marker scaffolds or haplotypes.

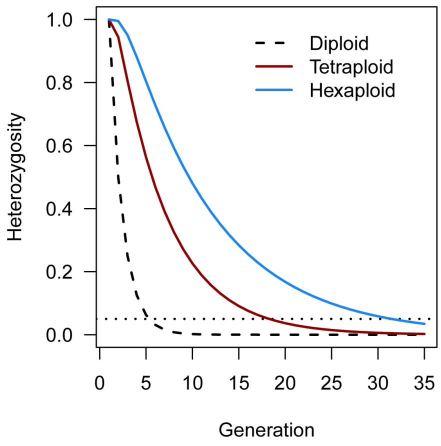


FIGURE 4 | Theoretical rate of decrease in heterozygosity in polyploid species from repeated rounds of inbreeding/selfing, using expressions derived by Haldane (1930). For autotetraploids (red line), 95% homozygosity (horizontal dotted line) is achieved after on average 19 generations of selfing, while for a hexaploid (blue line) 95% homozygosity is reached after approximately 32 generations. By contrast, a diploid reaches 95% homozygosity after approximately 5 generations of selfing (black dashed line).

These are then used to calculate recombination frequencies by counting recombination events both within and between these scaffolds, leading to an extremely simple estimate of *r* which has no corresponding LOD score. Scaffolds are clustered using a graph partitioning algorithm, and thereafter, the computationally efficient CONCORDE traveling-salesman solver is employed to order markers [as is done for example in TSPmap (Monroe et al., 2017)]. This assumes that the variance of all *r* estimates is equal and that weights are not required – which may well be the case if the haplotype scaffolds are correctly constructed. PolyGembler claims to be able to handle the high levels of missing data and genotyping errors associated with next-generation sequencing data. Although it is applicable to multiple ploidy levels, the authors point out that mapping at the hexaploid level becomes computationally difficult due to the huge number of possible combinations in the formation of haplotypes. However, it appears to be a very promising tool which combines both genetic and bioinformatic approaches in a single pipeline.

Apart from those tools which constitute a complete linkage mapping pipeline, there have been some specific tools recently developed which we predict will have an important impact on future polyploid mapping applications. One of the most significant of these is the MDSMap package in R (Preedy and Hackett, 2016), a novel approach for determining a map order using multi-dimensional scaling. Marker data in polyploid species possesses variable information content, a fact that can be appreciated by considering the haplotype origin of markers of dosage 1 from a duplex marker in a tetraploid species. Certain combinations of markers provide very unambiguous information about co-inheritance, whereas others do not. Therefore, weights are required to prevent imprecise combinations from exerting a large influence on the map order. Before MDSMap was developed, the only reliable algorithm for ordering weighted recombination frequencies was the weighted regression algorithm from the original JoinMap implementation (Stam, 1993; Van Ooijen, 2006). However, this has the disadvantage of being very slow for higher numbers of marker and is therefore of limited use with current high-density marker datasets. The MDSMap approach can achieve similar results in a fraction of the time, and takes as its input the same information as JoinMap does, the pairwise recombination frequency estimates and logarithm of odds (LOD) scores, making this tool suitable for linkage map construction at any ploidy level, provided pairwise linkage analysis can be performed.

One final tool that has also proven useful for polyploid linkage map construction is the LPmerge package in R (Endelman and Plomion, 2014). LPmerge uses linear programming to remove the minimum number of constraints in marker order in order to create a conflict-free consensus map. It was originally developed to create integrated genetic maps from multiple (diploid) populations. That said, polyploids contain multiple copies of each chromosome and therefore also present a similar challenge if we consider each homolog map as originating from a different population, with non-simplex markers as bridging markers (mapped in more than one population). Homolog-specific maps are still regularly generated in polyploid mapping studies [e.g., in potato (Bourke et al., 2015, 2016), rose (Vukosavljev et al., 2016) or sweet potato (Shirasawa et al., 2017)], for which LPmerge (or a similarly efficient integration algorithm) could then be used to generate chromosomally integrated maps.

Physical Maps

Arguably, one of the most important “tools” in current genomics studies is access to a high-quality reference genome assembly. Species for which a reference genome assembly exists have even been classified as “model organisms” (Seeb et al., 2011), such is the importance and impact a genome can bring to research on that species. Without a reference sequence available, the scope of genomic research remains limited. For example, GWAS rely on knowledge of the relative position of SNP markers (usually on a physical map), and many sequencing applications rely on a reference assembly on which to map reads. A reference genome also facilitates the development of molecular markers (e.g., primer development), the comparison of results between different genetic studies (by providing a single reference map), as well as allowing comparisons of specific sequences such as genes, enabling prediction of gene function across related species.

Polyploid genomes are by definition more complex than diploid genomes, having multiple copies of each homologous chromosome. Many polyploid species are also outbreeding, leading to increased heterozygosity which is problematic in *de novo* assemblies and necessitates specialized approaches (Kajitani et al., 2014). The most common solution until now has been to sequence a representative diploid species. For example in highly heterozygous autotetraploid potato, a completely homozygous doubled monoploid (*S. tuberosum* group *Phureja* DM1-3) was sequenced (Potato Genome Sequencing Consortium, 2011) which still represents the primary reference sequence today[3].

[3] http://solanaceae.plantbiology.msu.edu/

In the case of allopolyploids, multiple diploid progenitor species are often sequenced instead [e.g., peanut (Bertioli et al., 2016)]. The emergence of the pan-genome concept, originally proposed for microbial species (Tettelin et al., 2005), has interesting implications for how highly heterozygous polyploid genomes will be presented in future. We have already mentioned the arrival of phased genomics with the sweet potato genome, which aimed to generate six chromosome-length phased assemblies for each of its 15 chromosomes (Yang et al., 2017a). In future, both pan-genomes and phased genomes are likely to play a bigger role in polyploid reference genomics. Examples of polyploid species that have so far been "sequenced" are listed in **Table 1**. This is by no means an exhaustive list, nor does it describe all developments for the listed species. For example, the sequence of allotetraploid *Coffea arabica* (which accounts for roughly 70% of all coffee production) has recently been assembled, with a draft assembly (*C. arabica* UCDv0.5) available on the Phytozome database[4]. What **Table 1** highlights is that at the time of writing, there were already a wide range of polyploid crop species that have well-developed genomic resources, despite the fact that in many cases these are from closely related or progenitor diploid species. In time, just like for coffee, we predict that direct sequencing of polyploid species themselves will gradually replace the haploidised reference sequences in importance and application, leading to more insights of direct relevance to polyploids.

QUANTITATIVE TRAIT ANALYSIS AND GENOMIC SELECTION

One of the main goals of genetic studies is to find causative associations between DNA polymorphisms and phenotypic traits. In domesticated species in particular, these studies are often performed with a practical aim: to develop marker-based methods of selecting superior lines in a breeding program. Traditional approaches such as bi-parental QTL mapping have been complemented in recent years by new methodologies such as GWAS and genomic selection. However, all these approaches require polyploid-specific solutions which can capture the increased complexity of polysomic inheritance. We look at the three most commonly used approaches for identifying quantitative trait variation and how specific software tools are helping to revolutionize polyploid plant breeding programs.

QTL Analysis

The term "QTL analysis" usually refers to studies that aim to detect regions of the genome [so-called quantitative trait loci (Geldermann, 1975)] that have a significant statistical association with a trait in specifically constructed experimental populations. These populations are most often created by crossing two contrasting parental lines ("bi-parental" populations), although there is increasing interest in using more complex population designs in order to increase the range of alleles and genetic backgrounds being studied [e.g., "MAGIC" populations (Huang et al., 2015)]. As already discussed, there is great difficulty in developing inbred lines by repeatedly selfing polyploids due to the sampling of alleles during polyploid gamete formation [in a diploid this sampling generates $\binom{2}{1}$ =2 combinations; for a tetraploid this rises to $\binom{4}{2}$ =6 and in a hexaploid $\binom{6}{3}$ =20 combinations, resulting in protracted heterozygosity (**Figure 4**)], not to mention the problem of inbreeding depression associated with many outcrossing polyploid species. Therefore, most QTL analyses in polyploid species have been performed using the directly segregating F_1 progeny of a cross between heterozygous parents (a "full sib" population). This leads to poor resolution of QTL positions when compared to the more popular diploid inbred populations like RILs etc., as well as the fact that populations must be vegetatively propagated if replication over years or different growing environments is desired. For many polyploid species, vegetative propagation is indeed possible (Herben et al., 2017) and F_1 populations have the added advantage of being relatively quick and simple to develop, while, because of a generally high level of heterozygosity, many loci will be segregating in the F_1. Therefore despite their drawbacks, F_1 populations remain the bi-parental population of choice for mapping studies.

The methods for QTL analysis in diploid species have become increasingly convoluted (van Eeuwijk et al., 2010); in polyploid species such theoretical complexities have yet to be attempted, given the more immediate difficulties in accurately genotyping as well as modeling polyploid inheritance. Just like for linkage mapping and GWAS, the range of software tools available for QTL analysis in polyploids remains rather limited, although there are a number of recent developments that are helping transform the field.

One of the only dedicated software for tetraploid QTL analysis is the already-mentioned TetraploidMap software (Hackett et al., 2007). This software enables interval mapping to be performed in autotetraploid F_1 populations (as well as a simple single-marker ANOVA test), using a restricted range of markers (1 × 0, 2 × 0, and 1 × 1 markers only, where 1 × 0 denotes a marker dosage of 1 in one parent and 0 in the other, etc.). Although still available, it has been superseded by the TetraploidSNPMap software (Hackett et al., 2017). TetraploidSNPMap (TSNPM) uses SNP dosage data to either construct a linkage map (as already described) or perform QTL interval mapping. In contrast to its predecessor, TSNPM can analyze all marker segregation types, and allows the user to explore different QTL models at detected peaks. At its core is an algorithm to determine identity-by-descent (IBD) probabilities for the offspring of the population, which are then used in a weighted regression performed across the genome.

An independent software tool that has been developed to determine IBD probabilities in tetraploids is TetraOrigin (Zheng et al., 2016), implemented in the Mathematica programming language. TetraOrigin relaxes the assumption of random bivalent pairing during meiosis (which TSNPM employs) to allow for

[4] www.phytozome.net

TABLE 1 | Some examples of publicly available reference sequences for polyploid species.

Target species	Sequenced species (ploidy)	Genome browser	Reference
Autopolyploids			
Alfalfa, *Medicago sativa* (4×)	*Medicago truncatula* (2×)	medicagogenome.org | plants.ensembl.org	Young et al., 2011; Tang et al., 2014
Kiwifruit, *Actinidia chinensis* (6×)	*Actinidia chinensis* (2×)	bdg.hfut.edu.cn/kir | bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi	Huang et al., 2013
Potato, *Solanum tuberosum* (4×)	*Solanum tuberosum* (2×)	solanaceae.plantbiology.msu.edu | plants.ensembl.org	Potato Genome Sequencing Consortium, 2011
Sweet potato, *Ipomoea batatas* (6×)	*Ipomoea batatas* (6×)	public-genomes-ngs.molgen.mpg.de/SweetPotato | ipomoea-genome.org	Yang et al., 2017a
Rose, *Rosa* × *hybrida* (4×)	*Rosa chinensis* (2×)	https://iris.angers.inra.fr/obh/	Hibrand-Saint Oyant et al., unpublished
Allopolyploids			
Banana, *Musa acuminata* (3×)	*Musa acuminata* (2×)	banana-genome-hub.southgreen.fr | plants.ensembl.org	D'Hont et al., 2012
Coffee, *Coffea arabica* (4×)	*Coffea canephora* (2×)	coffee-genome.org	Denoeud et al., 2014
Cotton, *Gossypium hirsutum* (4×)	*Gossypium hirsutum* (4×)	cottongen.org	Li et al., 2015
Oilseed rape, *Brassica napus* (4×)	*Brassica napus* (4×)	genoscope.cns.fr/brassicanapus | plants.ensembl.org	Chalhoub et al., 2014
Peanut, *Arachis hypogaea* (4×)	*Arachis duranensis* (2×) *Arachis ipaensis* (2×)	peanutbase.org	Bertioli et al., 2016
Quinoa, *Chenopodium quinoa* (4×)	*Chenopodium quinoa* (4×)	cbrc.kaust.edu.sa/chenopodiumdb	Jarvis et al., 2017
Strawberry, *Fragaria* × *ananassa* (8×)	*Fragaria vesca* (2×)	rosaceae.org	Shulaev et al., 2011
Wheat, *Triticum aestivum* (6×)	*Triticum aestivum* (6×)	wheat-urgi.versailles.inra.fr | plants.ensembl.org	International Wheat Genome Sequencing Consortium, 2014

both preferential chromosomal pairing as well as multivalent formation and the possibility of double reduction. Although not programmed in a user-friendly format like TSNPM, it is relatively straightforward to use, taking an integrated linkage map and marker dosage matrix as input. It does not perform QTL analysis directly, but the resulting IBD probabilities can then be used to model genotype effects in a QTL scan either using a weighted regression approach like TSNPM, or in a linear mixed model setting. IBD probabilities allow interval mapping since they can be interpolated at any desired intervals on the linkage map.

For ploidy levels other than tetraploid, there are currently no dedicated software tools available for QTL analysis or IBD probability estimation. Single-marker approaches such as ANOVA on the marker dosages [assuming additivity – various dominant models could also be explored; see, e.g., (Rosyara et al., 2016)] are of course possible and require access to basic statistical software packages such as R (or even Excel). However, such approaches are not ideal – they are only effective if marker alleles are closely linked in coupling with QTL alleles, and offer no ability to predict the QTL segregation type or mode of gene action as is done for example in TSNPM (Hackett et al., 2017). As interest increases in the genetic dissection of important traits in polyploid species, we anticipate that it is only a matter of time before more flexible cross-ploidy solutions are developed. Methodologies developed for tetraploid species often claim that "extension to higher ploidy levels is straightforward." These sorts of disingenuous claims attempt to mark new research territory as already solved. If extensions to higher ploidy levels were indeed straightforward we would already be reporting on a wider range of tools available for them – as far as we can tell, so far there are none.

Returning to the topic of population types, we also anticipate that more powerful QTL analyses can be performed by combining information over multiple populations. Approaches such as pedigree-informed analyses, implemented for diploids in the FlexQTL software (Bink et al., 2008), could overcome some of the limitations imposed by the restrictions on population types in software for polyploids. However, it may take some time before such tools become translated to the polyploid level.

Genome-Wide Association Studies

Genome-wide association studies have emerged as a powerful tool for detecting causative loci underlying phenotypic traits. They have been particularly popular in species where the generation of experimental populations is problematic (such as humans). GWAS has been readily adopted across a broad spectrum of species since then, due to the promise of increased mapping resolution, a more diverse sampling of alleles and a simplicity in population creation (no crossing required) (Bernardo, 2016). There are certain disadvantages though, particularly in how rare (and potentially important) variants can be missed (Ott et al., 2015) and the confounding effect of population structure on results (Korte and Farlow, 2013). Nevertheless, GWAS continues to be an important analytical option to help shed greater light on genotype – phenotype associations. The application of GWAS in polyploid species is relatively new, although there have already been a number of studies published in various crop species, for example in potato, oilseed rape, wheat, and oats (Uitdewilligen et al., 2013; Gajardo et al., 2015; Sukumaran et al., 2015; Tumino et al., 2016, 2017). GWAS studies usually need to account for population structure and relatedness to prevent spurious associations, often in the context of linear mixed models (Yu et al., 2006; Bradbury et al., 2007; Zhang et al., 2010).

One challenge in applying GWAS to polyploid species is how to define a relatedness metric between polyploid individuals (i.e., how to generate the kinship matrix, *K*). So far, there have been two software tools released for polyploid GWAS, namely the R package GWASpoly (Rosyara et al., 2016) and the previously mentioned SHEsisPlus (Shen et al., 2016). Of these, only GWASpoly looks critically at the form of the kinship matrix *K*. Three different forms of *K* were tested in the development of the package, with the canonical relationship matrix (VanRaden, 2008) [termed the realized relationship matrix by the authors (Rosyara et al., 2016)] found to best control against inflation of significance values. This is also the default *K* provided in the GWASpoly package. An alternative approach to GWAS mapping for polyploids is provided by the netgwas package (Behrouzi and Wit, 2017b), previously mentioned for its linkage mapping capacity. Again, graphical models form the basis of the approach, which goes beyond single-marker association mapping to investigate genotype-phenotype interactions using all markers simultaneously in a graph structure. There is almost no discussion on how confoundedness between population structure and phenotypes are handled, but the authors claim the detection of false positive associations is not problematic.

One final aspect worth considering is the issue of deploying an adequate number of markers in a polyploid GWAS, which potentially represents a much larger genomic space. In *A. thaliana*, it was estimated that between 140K and 250K SNPs would be needed to fully cover the genome based on a study of linkage disequilibrium in that species (Kim et al., 2007). Modeling the decay of linkage disequilibrium in polyploid species is a more complex exercise. It was previously suggested that estimates of linkage disequilibrium may be inflated in polyploid species (Jannoo et al., 1999; Flint-Garcia et al., 2003). A more recent survey of linkage disequilibrium in autotetraploid potato using SNP dosages estimated that at most 40K SNPs would be needed for QTL discovery in potato (Vos et al., 2017), a much lower estimate than for *Arabidopsis* (Kim et al., 2007). The discrepancy comes in part from the differences in how these figures were estimated, using a 'hide-the-SNP' simulation for *Arabidopsis* versus a 'rule of thumb' calculation for potato, but mainly from the difference in the extent of LD between the two species [estimated at ∼10 Kb in *A. thaliana* versus ∼2 Mb in *S. tuberosum* (Kim et al., 2007; Vos et al., 2017)]. Detecting or even defining linkage disequilibrium between markers linked in repulsion phase is non-trivial in autopolyploids (Vos et al., 2017), which is analogous to the problem of detecting and estimating recombination frequency between such markers in a linkage mapping study. So far, we are not aware of any software tool that has been developed to estimate the extent of linkage disequilibrium in polyploids, which would complement the design of future GWAS studies in polyploid species.

Genomic Prediction and Genomic Selection

There has been much attention given to the advantages of using *all* marker data to help predict phenotypic performance, rather than focussing on single markers (or haplotypes) that are linked to QTL as was previously advocated. The motivation behind this is clear – many of the most important traits in domesticated animal and plant species are highly quantitative, with far too many small-effect loci present to be able to tag them all with single markers (Bernardo, 2008). One of the most important traits in any breeding program is also a famously quantitative trait: yield. It has been suggested that despite many years of phenotypic selection, crop yield in tetraploid potato has essentially remained unchanged (Jansky, 2009; Slater et al., 2016). This is a remarkable indictment of traditional selection methods, yet offers much-needed impetus for the development and deployment of new paradigms in breeding for quantitative traits.

Genomic prediction first arose in animal breeding circles (Meuwissen et al., 2001), where the concept of estimating breeding values from known pedigrees was already well-established. However, the estimation of breeding values in polyploid species requires special consideration due to the complexity of polysomic inheritance and the possibility of double reduction. In practice, breeding values are usually estimated using restricted maximum likelihood (REML) to solve mixed model equations, requiring the generation of an inverse additive relationship matrix A^{-1}, also called the numerator relationship matrix. The form of A^{-1} depends on, among other things, whether the inheritance is polysomic or disomic, and whether double reduction occurs (Kerr et al., 2012; Amadeu et al., 2016; Hamilton and Kerr, 2017). The R package AGHmatrix was developed in order to compute the appropriate *A* matrix for autotetraploids with a known pedigree (Amadeu et al., 2016), using theory developed in (Kerr et al., 2012). In applying their approach to an autotetraploid blueberry (*Vaccinium corymbosum* L.) population, the authors determined the *A* matrix under various levels of double reduction, afterwards selecting the model which maximized the likelihood of the data (Amadeu et al., 2016). More recently, an alternative R package polyAinv was released which computes A^{-1} as well as the kinship matrix *K* and the inbreeding coefficients *F* (Hamilton and Kerr, 2017). polyAinv claims to be applicable to any ploidy level (rather than just autotetraploids) and can accommodate sex-based differences in IBD probabilities (Hamilton and Kerr, 2017). Like AGHmatrix, it also incorporates double reduction in its calculations. However, in one study of nine common traits in autotetraploid potato, the inclusion of double reduction, or even the adoption of an autotetraploid-appropriate relationship matrix was found to have a minimal impact on the results (Slater et al., 2014). Studies which ignore the specific complexities of autopolyploids may still benefit from genomic prediction and selection, as for example was demonstrated in tetraploid potato (Sverrisdóttir et al., 2017).

Commonly used software tools for estimating breeding values at the diploid level include ProGeno (Maenhout, 2018) and ASreml (VSN International, 2018) which could be suitable for polyploid breeding programs, although this has yet to be conclusively demonstrated.

POLYPLOID INHERITANCE AND SIMULATION

As a final section we look at two topics which are important to the development of polyploid genetic resources – the mode of inheritance and the availability of simulation software for polyploid species. Although these topics do not necessarily go together, they represent very important considerations in themselves. The mode of inheritance is a polyploid-specific topic, with no equivalent issue arising in diploid genetic studies. Simulation studies, on the other hand, have been used repeatedly at the diploid level to test new methodologies, determine empirical thresholds, evaluate competing methods etc. The availability of a range of software options to simulate polyploid genetic behavior is crucial if polyploid genetics is to flourish.

Mode of Inheritance

The term “mode of inheritance” refers to the randomness of meiotic pairing processes that give rise to gametes, and is often used to distinguish between disomic (diploid-like) inheritance, and polysomic (all allele combinations equally possible) inheritance. As alluded to already, intermediate modes of inheritance are theoretically possible if partially preferential pairing occurs between homologs, resulting in on average more recombinations between certain homologs, and less between others (putative homoeologs). This intermediate inheritance pattern, originally termed segmental allopolyploidy (Stebbins, 1947) and more recently termed mixosomy (Soltis et al., 2016), poses additional challenges over those of purely polysomic or disomic behavior. One of the main complications is the lack of fixed segregation ratios to test markers against (Allendorf and Danzmann, 1997), which is often used as a measure of marker quality (Stringham and Boehnke, 1996; Pompanon et al., 2005). Currently there are no dedicated tools available to ascertain the most likely mode of inheritance in polyploids. Some “traditional” approaches to predict the mode of inheritance are summarized in (Bourke et al., 2017), many of which are relatively straightforward to implement using a statistical programming environment like R (R Core Team, 2016). In that study, TetraOrigin (Zheng et al., 2016) was used to estimate the most likely pairing configuration that gave rise to each offspring in an F_1 tetraploid population. This enabled the authors to test whether there were deviations from the expected patterns of homolog pairing under a tetrasomic model (Bourke et al., 2017). A simple alternative using closely linked repulsion-phase simplex marker pairs was also proposed and has been implemented in the polymapR package (Bourke et al., unpublished). Apart from preferential pairing, TetraOrigin can also predict whether marker data arose from bivalent or multivalent pairing during meiosis, facilitating an analysis of the distribution of double reduction products. However, apart from its restriction to tetraploid data, an integrated linkage map is required before TetraOrigin can be employed. In severe cases of mixosomy, it is not obvious how a reliable linkage map should be generated. Corrections for mixosomy in a tetraploid linkage analysis are possible in polymapR, but in extreme cases marker clustering will also be affected, making map construction quite challenging. A confounding complication is the possibility of variable chromosome counts (aneuploidy), as for example encountered in sugarcane (Grivet et al., 1996; Grivet and Arruda, 2002) or in ornamentals such as *Alstroemeria* (Buitendijk et al., 1997), which makes the diagnosis of the mode of inheritance even more difficult. As more polyploid species begin to be genotyped, the issue of unknown mode of inheritance will likely exert more influence, further necessitating the development of software tools that can provide an accurate assessment of the inheritance mode using marker data, and that can accommodate the full spectrum of polyploid meiotic behaviors.

Simulation Software

As with any software tool, developing standards and scenarios upon which the performance of the tool can be judged is vital to ensure reliable results. In this final section we consider the range of simulation tools currently available for polyploids. Probably the most widely used polyploid simulation software currently available is PedigreeSim (Voorrips and Maliepaard, 2012). Originally developed to generate diploid and tetraploid populations, the current release (PedigreeSim V2.0) can simulate populations of any even ploidy level (2, 4, 6, . . .). What makes PedigreeSim particularly attractive is its ability to simulate a diversity of meiotic pairing conditions, including quadrivalents (which can result in double reduction) or preferential chromosome pairing. It takes four input files (which are relatively simple to generate) that provide a description of the desired simulation parameters and the input marker data. The software then creates (dosage-scored) genotype data for any pedigreed population, e.g., an F_1 population of specified size (Voorrips and Maliepaard, 2012). Some authors have used PedigreeSim to simulate multiple generations of random mating, allowing an investigation of population structure and linkage disequilibrium in polyploid species (e.g., Rosyara et al., 2016; Vos et al., 2017), which can be implemented quite easily with some basic programming knowledge. PedigreeSim is written in Java and can run on all major operating systems.

A Windows-based software Polylink, which originally performed two-point linkage analysis and simulation of tetraploid populations (He et al., 2001), is no longer available. The R package polySegratio (Baker, 2014) simulates dominantly scored marker data in autopolyploids of any even ploidy level. Generating the dosage data is straightforward: only the expected proportion of marker types (simplex, duplex, triplex, . . .) as well as the ploidy is required. However, the markers are essentially completely random, with no connection to any linkage map, which is arguably of limited use for any application that requires some degree of linkage between markers. The simulation capacities of polysegRatio therefore appear to be most useful for testing functions within the package itself, namely those designed to impute parental dosages given the observed segregation ratios in offspring scores.

A final polyploid simulation tool that has recently been developed is the HaploSim pipeline which includes the HaploGenerator function (Motazedi et al., 2017). HaploGenerator is designed to generate sequence-based haplotypes in a polyploid of any even ploidy, taking the fasta file it is provided with as a reference from which haplotypes are built. The software generates random SNP mutations at a specified distribution before simulating next-generation sequencing (NGS) reads in formats corresponding to a number of current sequencing technologies such as Illumina or Pacific Biosystems (PacBio). The pipeline was originally developed to compare the performance of a number of haplotype assembly algorithms (Motazedi et al., 2017), but could also be useful for testing the performance of any other tool which uses NGS reads as genotypes.

FUTURE PERSPECTIVES

In this review we have attempted to describe the most important software tools that are currently available to the polyploid genetics community. There are likely to be tools that were missed and tools that have subsequently been released – this is the danger of such a review. However, we have tried where possible to also discuss the gaps that are apparent in the current set of available tools which will hopefully help guide their development in future. Polyploid genotyping arguably remains the most critical step, as without accurate genotype data there is little point in building models for polyploid inheritance. However, we are now witnessing the slow emergence of tools that take polyploid genotypes and use them to make inferences on the transmission of alleles and the effects of such alleles in polyploid populations. As genotyping technologies continue to evolve, so too should the suite of tools developed to analyze those genotypes. Tools for analyzing SNP dosage data from SNP arrays are well-established. The coming decade will likely see a move away from SNP array-based genotyping to the use of sequence-read based genotypes, although this will require that all tools heretofore developed be updated to accommodate the new type of data. Information on the mode of inheritance from marker data is also needed for each population studied, which deserves more attention than it currently receives. A move from diploid-based reference genomes to fully polyploid (and haplotype-resolved) reference genomes would also help broaden the boundaries of polyploid genetics away from the diplo-centric view of genomics which currently dominates. Although there have been many exciting discoveries and developments in polyploid genetics in the past decade or more, we feel its golden age has yet to arrive, an age which will be heralded all the sooner by the provision of robust and user-friendly tools for the genetic dissection of this fascinating group of organisms.

AUTHOR CONTRIBUTIONS

PB wrote this review, with input from REV, RGFV, and CM. All authors read and approved the final manuscript.

FUNDING

This research was provided through the TKI polyploids project "Novel Genetic and Genomic Tools for Polyploid Crops" (project numbers BO-26.03-009-004 and BO-50-002-022). The support of the companies participating in the polyploids projects is gratefully acknowledged.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Jeffrey Endelman (University of Wisconsin–Madison) for the helpful remarks regarding tetraploid kinship matrices, and Dr. Heleen Bastiaanssen (Anthura B.V.) for the helpful feedback on the manuscript.

REFERENCES

Aguiar, D., and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* 29, i352–i360. doi: 10.1093/bioinformatics/btt213

Akhunov, E., Nicolet, C., and Dvorak, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor. Appl. Genet.* 119, 507–517. doi: 10.1007/s00122-009-1059-5

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi: 10.1093/molbev/msq148

Allendorf, F. W., and Danzmann, R. G. (1997). Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* 145, 1083–1092.

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A., Resende, M. F., and Muñoz, P. R. (2016). AGHmatrix: R Package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2016.01.0009

Bajgain, P., Rouse, M. N., and Anderson, J. A. (2016). Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop Sci.* 56, 232–248. doi: 10.2135/cropsci2015.06.0389

Baker, P. (2014). polySegratio: simulate and test marker dosage for dominant markers in autopolyploids. R Package Version 0.2–4. doi: 10.1007/s00122-010-1283-z

Baker, P., Jackson, P., and Aitken, K. (2010). Bayesian estimation of marker dosage in sugarcane and other autopolyploids. *Theor. Appl. Genet.* 120, 1653–1672. doi: 10.1007/s00122-010-1283-z

Balsalobre, T. W. A., Da Silva Pereira, G., Margarido, G. R. A., Gazaffi, R., Barreto, F. Z., Anoni, C. O., et al. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics* 18:72. doi: 10.1186/s12864-016-3383-x

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Barringer, B. C. (2007). Polyploidy and self-fertilization in flowering plants. *Am. J. Bot.* 94, 1527–1533. doi: 10.3732/ajb.94.9.1527

Bartholomé, J., Mandrou, E., Mabiala, A., Jenkins, J., Nabihoudine, I., Klopp, C., et al. (2015). High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytol.* 206, 1283–1296. doi: 10.1111/nph.13150

Bassil, N. V., Davis, T. M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., et al. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16:155. doi: 10.1186/s12864-015-1310-1

Behrouzi, P., and Wit, E. C. (2017a). De novo construction of q-ploid linkage maps using discrete graphical models. arXiv preprint arXiv:1710.01063.

Behrouzi, P., and Wit, E. C. (2017b). netgwas: an R package for network-based genome-wide association studies. arXiv preprint arXiv:1710.01236.

Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). Haptree: A novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput. Biol.* 10:e1003502. doi: 10.1371/journal.pcbi.1003502

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48, 1649–1664. doi: 10.2135/cropsci2008.03.0131

Bernardo, R. (2016). Bandwagons I, too, have known. *Theor. Appl. Genet.* 129, 2323–2332. doi: 10.1007/s00122-016-2772-5

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

Bink, M., Boer, M., Ter Braak, C., Jansen, J., Voorrips, R., and Van De Weg, W. (2008). Bayesian analysis of complex traits in pedigreed plant populations. *Euphytica* 161, 85–96. doi: 10.1007/s10681-007-9516-1

Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F. S., Van 't Westende, W. P. C., et al. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *Plant J.* 90, 330–343. doi: 10.1111/tpj.13496

Bourke, P. M., Voorrips, R. E., Kranenburg, T., Jansen, J., Visser, R. G., and Maliepaard, C. (2016). Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theor. Appl. Genet.* 129, 2211–2226. doi: 10.1007/s00122-016-2768-1

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2015). The double reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics* 201, 853–863. doi: 10.1534/genetics.115.181008

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

Buitendijk, J. H., Boon, E. J., and Ramanna, M. S. (1997). Nuclear DNA content in twelve species of *Alstroemeria* L. and some of their hybrids. *Ann. Bot.* 79, 343–353. doi: 10.1006/anbo.1996.0345

Bushman, B., Robbins, M., Larson, S., and Staub, J. (eds) (2016). "Genotyping by sequencing in autotetraploid cocksfoot (*Dactylis glomerata*) without a reference genome," in *Breeding in a World of Scarcity*, (Berlin: Springer), 133–137. doi: 10.1007/978-3-319-28932-8_20

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

Chaffin, A. S., Huang, Y.-F., Smith, S., Bekele, W. A., Babiker, E., Gnanesh, B. N., et al. (2016). A consensus map in cultivated hexaploid oat reveals conserved grass synteny with substantial subgenome rearrangement. *Plant Genome* 9, 1–21. doi: 10.3835/plantgenome2015.10.0102

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Cheema, J., and Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Brief. Bioinform.* 10, 595–608. doi: 10.1093/bib/bbp045

Chester, M., Gallagher, J. P., Symonds, V. V., Da Silva, A. V. C., Mavrodiev, E. V., Leitch, A. R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 109, 1176–1181. doi: 10.1073/pnas.1112041109

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* 129, 1887–1899. doi: 10.1007/s00122-016-2746-7

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., et al. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Mol. Ecol.* 25, 616–629. doi: 10.1111/mec.13491

Crawford, J., Brown, P. J., Voigt, T., and Lee, D. (2016). Linkage mapping in prairie cordgrass (*Spartina pectinata* Link) using genotyping-by-sequencing. *Mol. Breed.* 36:62. doi: 10.1007/s11032-016-0484-9

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., et al. (2014). A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Funct. Integr. Genomics* 14, 643–655. doi: 10.1007/s10142-014-0391-2

Das, S., and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* 16:260. doi: 10.1186/s12864-015-1408-5

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223. doi: 10.1038/ng1669

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345, 1181–1184. doi: 10.1126/science.1255274

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, 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

Didion, J. P., Yang, H., Sheppard, K., Fu, C.-P., Mcmillan, L., De Villena, F. P.-M., et al. (2012). Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 13:34. doi: 10.1186/1471-2164-13-34

Doyle, J. J., and Sherman-Broyles, S. (2016). Double trouble: taxonomy and definitions of polyploidy. *New Phytol.* 213, 487–493. doi: 10.1111/nph.14276

Edae, E. A., Bowden, R. L., and Poland, J. (2015). Application of population sequencing (POPSEQ) for ordering and imputing genotyping-by-sequencing markers in hexaploid wheat. *G3* 5, 2547–2553. doi: 10.1534/g3.115.020362

Endelman, J. B., and Plomion, C. (2014). LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* 30, 1623–1624. doi: 10.1093/bioinformatics/btu091

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., et al. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One* 7:e36347. doi: 10.1371/journal.pone.0036347

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* 6:220. doi: 10.3389/fgene.2015.00220

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. IV. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Gajardo, H. A., Wittkop, B., Soto-Cerda, B., Higgins, E. E., Parkin, I. A. P., Snowdon, R. J., et al. (2015). Association mapping of seed quality traits in *Brassica napus* L. using GWAS and candidate QTL approaches. *Mol. Breed.* 35:143. doi: 10.1007/s11032-015-0340-3

Galloway, L. F., and Etterson, J. R. (2007). Inbreeding depression in an autotetraploid herb: a three cohort field study. *New Phytol.* 173, 383–392. doi: 10.1111/j.1469-8137.2006.01909.x

Galloway, L. F., Etterson, J. R., and Hamrick, J. L. (2003). Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana. Heredity* 90, 308–315. doi: 10.1038/sj.hdy.6800242

Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *TAG Theor. Appl. Genet.* 46, 319–330. doi: 10.1007/BF00281673

Gidskehaug, L., Kent, M., Hayes, B. J., and Lien, S. (2010). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* 27, 303–310. doi: 10.1093/bioinformatics/btq673

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: what are they and how do we infer them? *Trends Plant Sci.* 21, 609–621. doi: 10.1016/j.tplants.2016.02.005

Goldschmidt, R. (1933). Some aspects of evolution. *Science* 78, 539–547. doi: 10.1126/science.78.2033.539

Grandke, F., Ranganathan, S., Van Bers, N., De Haan, J. R., and Metzler, D. (2017). PERGOLA: fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics* 18:12. doi: 10.1186/s12859-016-1416-8

Grivet, L., and Arruda, P. (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5, 122–127. doi: 10.1016/S1369-5266(02)00234-0

Grivet, L., D'hont, A., Roques, D., Feldmann, P., Lanaud, C., and Glaszmann, J. C. (1996). RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics* 142, 987–1000.

Hackett, C., and Luo, Z. (2003). TetraploidMap: construction of a linkage map in autotetraploid species. *J. Hered.* 94, 358–359. doi: 10.1093/jhered/esg066

Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F., and Milne, I. (2017). TetraploidSNPMap: Software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *J. Hered.* 108, 438–442. doi: 10.1093/jhered/esx022

Hackett, C. A., Mclean, K., and Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS One* 8:e63939. doi: 10.1371/journal.pone.0063939

Hackett, C. A., Milne, I., Bradshaw, J. E., and Luo, Z. (2007). TetraploidMap for windows: linkage map construction and QTL mapping in autotetraploid species. *J. Hered.* 98, 727–729. doi: 10.1093/jhered/esm086

Haldane, J. B. (1930). Theoretical genetics of autopolyploids. *J. Genet.* 22, 359–372. doi: 10.1007/BF02984197

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., et al. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* 12:302. doi: 10.1186/1471-2164-12-302

Hamilton, M. G., and Kerr, R. J. (2017). Computation of the inverse additive relationship matrix for autopolyploid and multiple-ploidy populations. *Theor. Appl. Genet.* 131, 851–860. doi: 10.1007/s00122-017-3041-y

Harlan, J. R., and De Wet, J. M. J. (1975). On Ö. Winge and a prayer: the origins of polyploidy. *Bot. Rev.* 41, 361–390. doi: 10.1007/BF02860830

He, Y., Xu, X., Tobutt, K. R., and Ridout, M. S. (2001). Polylink: to support two-point linkage analysis in autotetraploids. *Bioinformatics* 17, 740–741. doi: 10.1093/bioinformatics/17.8.740

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., et al. (2015). MAGIC populations in crops: current status and future prospects. *Theor. Appl. Genet.* 128, 999–1017. doi: 10.1007/s00122-015-2506-0

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., et al. (2013). Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* 4:2640. doi: 10.1038/ncomms3640

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63K SNP Array for cotton and high-density mapping of intra- and inter-specific populations of *Gossypium* spp. *G3* 5, 1187–1209. doi: 10.1534/g3.115.018416

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788. doi: 10.1126/science.1251788

Islam, M. S., Thyssen, G. N., Jenkins, J. N., and Fang, D. D. (2015). Detection, validation, and application of genotyping-by-sequencing based single nucleotide polymorphisms in Upland cotton. *Plant Genome* 8, 1–10. doi: 10.3835/plantgenome2014.07.0034

Jannoo, N., Grivet, L., Dookun, A., D'hont, A., and Glaszmann, J. C. (1999). Linkage disequilibrium among modern sugarcane cultivars. *Theor. Appl. Genet.* 99, 1053–1060.

Jansky, S. (2009). "Chapter 2 - Breeding, genetics, and cultivar development," in *Advances in Potato Chemistry and Technology*, eds J. Singh and L. Kaur (San Diego, CA: Academic Press), 27–62.

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., et al. (2017). The genome of *Chenopodium quinoa*. *Nature* 542, 307–312. doi: 10.1038/nature21370

Jones, D. B., Jerry, D. R., Khatkar, M. S., Raadsma, H. W., Van Der Steen, H., Prochaska, J., et al. (2017). A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, *Litopenaeus vannamei*. *Sci. Rep.* 7:10360. doi: 10.1038/s41598-017-10515-7

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi: 10.1101/gr.170720.113

Kantarski, T., Larson, S., Zhang, X., Dehaan, L., Borevitz, J., Anderson, J., et al. (2017). Development of the first consensus genetic map of intermediate wheatgrass (*Thinopyrum intermedium*) using genotyping-by-sequencing. *Theor. Appl. Genet.* 130, 137–150. doi: 10.1007/s00122-016-2799-7

Kaur, S., Francki, M. G., and Forster, J. W. (2012). Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnol. J.* 10, 125–138. doi: 10.1111/j.1467-7652.2011.00644.x

Kerr, R. J., Li, L., Tier, B., Dutkowski, G. W., and Mcrae, T. A. (2012). Use of the numerator relationship matrix in genetic analysis of autopolyploid species. *Theor. Appl. Genet.* 124, 1271–1282. doi: 10.1007/s00122-012-1785-y

Kihara, H., and Ono, T. (1926). Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Z. Zellforsch. Mikrosk. Anat.* 4, 475–481. doi: 10.1007/BF00391215

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., and Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci.* 242, 14–22. doi: 10.1016/j.plantsci.2015.04.016

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., et al. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 39, 1151–1155. doi: 10.1038/ng2115

Koning-Boucoiran, C. F. S., Esselink, G. D., Vukosavljev, M., Van't Westende, W. P. C., Gitonga, V. W., Krens, F. A., et al. (2015). Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (Rosa L.). *Front. Plant Sci.* 6:249. doi: 10.3389/fpls.2015.00249

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Krebs, S. L., and Hancock, J. F. (1990). Early-acting inbreeding depression and reproductive success in the highbush blueberry, *Vaccinium corymbosum* L. *Theor. Appl. Genet.* 79, 825–832. doi: 10.1007/BF00224252

Lachance, J., and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35, 780–786. doi: 10.1002/bies.201300014

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., et al. (2014). Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS One* 9:e84329. doi: 10.1371/journal.pone.0084329

Limborg, M. T., Seeb, L. W., and Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Mol. Ecol.* 25, 2117–2129. doi: 10.1111/mec.13601

Maenhout, S. (2018). *Progeno*. Gent: Ghent University.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118. doi: 10.1038/nmeth.1419

Mason, A. S. (2015). "Challenges of genotyping polyploid species," in *Plant Genotyping: Methods and Protocols*, ed. J. Batley (New York, NY: Springer), 161–168.

Mason, A. S., Higgins, E. E., Snowdon, R. J., Batley, J., Stein, A., Werner, C., et al. (2017). A user guide to the Brassica 60K Illumina Infinium^{TM} SNP genotyping array. *Theor. Appl. Genet.* 130, 621–633. doi: 10.1007/s00122-016-2849-1

Mather, K. (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *J. Genet.* 30, 53–78. doi: 10.1007/BF02982205

McAllister, C. A., and Miller, A. J. (2016). Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *Am. J. Bot.* 103, 1314–1325. doi: 10.3732/ajb.1600146

McCallum, S., Graham, J., Jorgensen, L., Rowland, L. J., Bassil, N. V., Hancock, J. F., et al. (2016). Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. *Mol. Breed.* 36:41. doi: 10.1007/s11032-016-0443-5

McCamy, P., Holloway, H., Yu, X., Dunne, J. C., Schwartz, B. M., Patton, A. J., et al. (2018). A SNP-based high-density linkage map of zoysiagrass (*Zoysia japonica* Steud.) and its use for the identification of QTL associated with winter hardiness. *Mol. Breed.* 38:10. doi: 10.1007/s11032-017-0763-0

Mendel, J. G. (1866). "Versuche über pflanzenhybriden," in *Verhandlungen des Naturforschenden Vereines in Brünn Bd*. IV Abhandlungen, 3–47.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Moncada, M. D. P., Tovar, E., Montoya, J. C., González, A., Spindel, J., and Mccouch, S. (2016). A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. *Tree Genet. Genomes* 12, 1–17. doi: 10.1007/s11295-015-0927-1

Monroe, J. G., Allen, Z. A., Tanger, P., Mullen, J. L., Lovell, J. T., Moyers, B. T., et al. (2017). TSPmap, a tool making use of traveling salesperson problem solvers in the efficient and accurate construction of high-density genetic linkage maps. *BioData Min.* 10:38. doi: 10.1186/s13040-017-0158-0

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A., and Russell, J. R. (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor. Appl. Genet.* 120, 1525–1534. doi: 10.1007/s00122-010-1273-1

Motazedi, E., Finkers, R., Maliepaard, C., and De Ridder, D. (2017). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Br. Bioinformat.* doi: 10.1093/bib/bbw126 [Epub ahead of print].

Neigenfind, J., Gyetvai, G., Basekow, R., Diehl, S., Achenbach, U., Gebhardt, C., et al. (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics* 9:356. doi: 10.1186/1471-2164-9-356

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276. doi: 10.1038/nature08250

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* 16, 275–284. doi: 10.1038/nrg3908

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 58K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* 7:40577. doi: 10.1038/srep40577

Peace, C., Bassil, N., Main, D., Ficklin, S., Rosyara, U. R., Stegmeir, T., et al. (2012). Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLoS One* 7:e48305. doi: 10.1371/journal.pone.0048305

Piepho, H.-P., and Koch, G. (2000). Codominant analysis of banding data from a dominant marker system by normal mixtures. *Genetics* 155, 1459–1468.

Plummer, M. (2003). "JAGS: A. (program) for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, 125.

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7:e32253. doi: 10.1371/journal.pone.0032253

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* 6, 847–859. doi: 10.1038/nrg1707

Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Preedy, K. F., and Hackett, C. A. (2016). A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theor. Appl. Genet.* 129, 2117–2132. doi: 10.1007/s00122-016-2761-8

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reddy, U. K., Nimmakayala, P., Abburi, V. L., Reddy, C., Saminathan, T., Percy, R. G., et al. (2017). Genome-wide divergence, haplotype distribution and population demographic histories for Gossypium hirsutum and *Gossypium barbadense* as revealed by genome-anchored SNPs. *Sci. Rep.* 7:41285. doi: 10.1038/srep41285

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2015.08.0073

Rothfels, C. J., Pryer, K. M., and Li, F. W. (2017). Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytol.* 213, 413–429. doi: 10.1111/nph.14111

Schmitz Carley, C. A., Coombs, J. J., Douches, D. S., Bethke, P. C., Palta, J. P., Novy, R. G., et al. (2017). Automated tetraploid genotype calling by hierarchical clustering. *Theor. Appl. Genet.* 130, 717–726. doi: 10.1007/s00122-016-2845-5

Seeb, J., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* 11, 1–8. doi: 10.1111/j.1755-0998.2010.02979.x

Serang, O., Mollinari, M., and Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS One* 7:e30906. doi: 10.1371/journal.pone.0030906

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., and Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Sci. Rep.* 6:24095. doi: 10.1038/srep24095

Shirasawa, K., Tanaka, M., Takahata, Y., Ma, D., Cao, Q., Liu, Q., et al. (2017). A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). *Sci. Rep.* 7:44207. doi: 10.1038/srep44207

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Slater, A. T., Cogan, N. O., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2016.02.0021

Slater, A. T., Wilson, G. M., Cogan, N. O., Forster, J. W., and Hayes, B. J. (2014). Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. *Theor. Appl. Genet.* 127, 809–820. doi: 10.1007/s00122-013-2258-7

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Soltis, P. S., and Soltis, D. E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 97, 7051–7057. doi: 10.1073/pnas.97.13.7051

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., et al. (2013). Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor. Appl. Genet.* 126, 2699–2716. doi: 10.1007/s00122-013-2166-x

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: join Map. *Plant J.* 3, 739–744. doi: 10.1111/j.1365-313X.1993.00739.x

Stebbins, G. L. (1947). Types of polyploids: their classification and significance. *Adv. Genet.* 1, 403–429.

Stringham, H. M., and Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis. *Am. J. Hum. Genet.* 59, 946–950.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool. Part A Ecol. Genet. Physiol.* 14, 43–59. doi: 10.1002/jez.1400140104

Su, S.-Y., White, J., Balding, D. J., and Coin, L. J. (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformat.* 9:513. doi: 10.1186/1471-2105-9-513

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., and Reynolds, M. P. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theor. Appl. Genet.* 128, 353–363. doi: 10.1007/s00122-014-2435-3

Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., et al. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet.* 130, 2091–2108. doi: 10.1007/s00122-017-2944-y

Swaminathan, M. S., and Howard, H. (1953). Cytology and genetics of the potato (*Solanum tuberosum*) and related species. *Bibliogr. Genet.* 16, 1–192.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4. 0) for the model legume *Medicago truncatula*. *BMC Genomics* 15:312. doi: 10.1186/1471-2164-15-312

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Tinker, N. A., Bekele, W. A., and Hattori, J. (2016). Haplotag: software for haplotype-based genotyping-by-sequencing analysis. *G3* 6, 857–863. doi: 10.1534/g3.115.024596

Tinker, N. A., Chao, S., Lazo, G. R., Oliver, R. E., Huang, Y.-F., Poland, J. A., et al. (2014). A SNP genotyping array for hexaploid oat. *Plant Genome* 7, 1–8. doi: 10.3835/plantgenome2015.10.0102

Tumino, G., Voorrips, R. E., Morcia, C., Ghizzoni, R., Germeier, C. U., Paulo, M.-J., et al. (2017). Genome-wide association analysis for lodging tolerance and plant height in a diverse European hexaploid oat collection. *Euphytica* 213:163. doi: 10.1007/s10681-017-1939-8

Tumino, G., Voorrips, R. E., Rizza, F., Badeck, F. W., Morcia, C., Ghizzoni, R., et al. (2016). Population structure and genome-wide association analysis for frost tolerance in oat using continuous SNP array signal intensity ratios. *Theor. Appl. Genet.* 129, 1711–1724. doi: 10.1007/s00122-016-2734-y

Uitdewilligen, J. G., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., and Van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One* 8:e62355. doi: 10.1371/journal.pone.0062355

van Dijk, T., Noordijk, Y., Dubos, T., Bink, M. C., Meulenbroek, B. J., Visser, R. G., et al. (2012). Microsatellite allele dose and configuration establishment (MADCE): an integrated approach for genetic studies in allopolyploids. *BMC Plant Biol.* 12:25. doi: 10.1186/1471-2229-12-25

van Eeuwijk, F. A., Bink, M. C., Chenu, K., and Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. *Curr. Opin. Plant Biol.* 13, 193–205. doi: 10.1016/j.pbi.2010.01.001

van Geest, G., Bourke, P. M., Voorrips, R. E., Marasek-Ciolakowska, A., Liao, Y., Post, A., et al. (2017a). An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theor. Appl. Genet.* 130, 2527–2541. doi: 10.1007/s00122-017-2974-5

van Geest, G., Voorrips, R. E., Esselink, D., Post, A., Visser, R. G., and Arens, P. (2017b). Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183 k SNP array. *BMC Genomics* 18:585. doi: 10.1186/s12864-017-4003-0

Van Ooijen, J. W. (2006). *JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen: Kyazma B.V.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Voorrips, R. E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12:172. doi: 10.1186/1471-2105-12-172

Voorrips, R. E., and Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics* 13:248. doi: 10.1186/1471-2105-13-248

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., Van Eck, H. J., and Van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130, 123–135. doi: 10.1007/s00122-016-2798-8

Vos, P. G., Uitdewilligen, J. G., Voorrips, R. E., Visser, R. G. F., and Van Eck, H. J. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* 128, 2387–2401. doi: 10.1007/s00122-015-2593-y

VSN International (2018). *ASreml*. Hempstead: VSN International Ltd.

Vukosavljev, M., Arens, P., Voorrips, R., Van 't Westende, W., Esselink, G. D., Bourke, P. M., et al. (2016). High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array. *Hortic. Res.* 3:16052. doi: 10.1038/hortres.2016.52

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* 14, 1195–1206. doi: 10.1111/pbi.12485

Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., et al. (2017a). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat. Plants* 3, 696–703. doi: 10.1038/s41477-017-0002-z

Yang, X., Sood, S., Glynn, N., Islam, M. S., Comstock, J., and Wang, J. (2017b). Constructing high-density genetic maps for polyploid sugarcane (*Saccharum* spp.) and identifying quantitative trait loci controlling brown rust resistance. *Mol. Breed.* 37:116. doi: 10.1007/s11032-017-0716-7

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* 9:104. doi: 10.3389/fpls.2018.00104

Young, N. D., Debellé, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 480, 520–524. doi: 10.1038/nature10625

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Yu, L. X., Zheng, P., Zhang, T., Rodringuez, J., and Main, D. (2017). Genotyping-by-sequencing-based genome-wide association studies on *Verticillium* wilt resistance in autotetraploid alfalfa (*Medicago sativa* L.). *Mol. Plant Pathol.* 18, 187–194. doi: 10.1111/mpp.12389

Zhang, K., Calabrese, P., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* 71, 1386–1394.

Zhang, T., Yu, L.-X., Zheng, P., Li, Y., Rivera, M., Main, D., et al. (2015). Identification of loci associated with drought resistance traits in heterozygous autotetraploid alfalfa (*Medicago sativa* L.) using genome-wide association studies with genotyping by sequencing. *PLoS One* 10:e0138931. doi: 10.1371/journal.pone.0138931

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., and Bink, M. C. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics* 203, 119–131. doi: 10.1534/genetics.115.185579

Conflict of Interest Statement: The authors of this review have been involved in the development of a number of the software tools mentioned, namely fitTetra, fitPoly, PedigreeSim, TetraOrigin, polymapR, HaploSim, and TriPoly. We have tried to give an impartial perspective where these and alternative tools are concerned. Apart from this, the authors do not have any further conflicts of interest to declare.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 24 July 2018
doi: 10.3389/fevo.2018.00107

Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context

Luís Felipe V. Ferrão[1†], *Juliana Benevenuto*[1†], *Ivone de Bem Oliveira*[1,2], *Catherine Cellon*[3], *James Olmstead*[4], *Matias Kirst*[5,6], *Marcio F. R. Resende Jr.*[7] *and Patricio Munoz*[1*]

[1] Blueberry Breeding and Genomics Laboratory, Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [2] Plant Genetics and Genomics Laboratory, Agronomy College, Federal University of Goias, Goiania, Brazil, [3] Duda Farm Fresh Foods, Oviedo, FL, United States, [4] Driscoll's Inc., Watsonville, CA, United States, [5] Forest Genomics Laboratory, School of Forest Resources and Conservation, University of Florida, Gainesville, FL, United States, [6] Genetics Institute, University of Florida, Gainesville, FL, United States, [7] Sweet Corn Genomics and Breeding, Horticultural Sciences Department, University of Florida, Gainesville, FL, United States

OPEN ACCESS

Edited by:
Hans D. Daetwyler,
La Trobe University, Australia

Reviewed by:
Peter Bradbury,
United States Department of Agriculture, United States
Ross Houston,
University of Edinburgh, United Kingdom

***Correspondence:**
Patricio Munoz
p.munoz@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 Ecology and Evolution

Received: *31 January 2018*
Accepted: *02 July 2018*
Published: *24 July 2018*

Citation:
Ferrão LFV, Benevenuto J, de Bem Oliveira I, Cellon C, Olmstead J, Kirst M, Resende MFR Jr and Munoz P (2018) Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context. Front. Ecol. Evol. 6:107. doi: 10.3389/fevo.2018.00107

Polyploidization is an ancient and recurrent process in plant evolution, impacting the diversification of natural populations and plant breeding strategies. Polyploidization occurs in many important crops; however, its effects on inheritance of many agronomic traits are still poorly understood compared with diploid species. Higher levels of allelic dosage or more complex interactions between alleles could affect the phenotype expression. Hence, the present study aimed to dissect the genetic basis of fruit-related traits in autotetraploid blueberries and identify candidate genes affecting phenotypic variation. We performed a genome-wide association study (GWAS) assuming diploid and tetraploid inheritance, encompassing distinct models of gene action (additive, general, different orders of allelic interaction, and the corresponding diploidized models). A total of 1,575 southern highbush blueberry individuals from a breeding population of 117 full-sib families were genotyped using sequence capture and next-generation sequencing, and evaluated for eight fruit-related traits. For the diploid allele calling, 77,496 SNPs were detected; while 80,591 SNPs were obtained in tetraploid, with a high degree of overlap (95%) between them. A linear mixed model that accounted for population and family structure was used for the GWAS analyses. By modeling tetraploid genotypes, we detected 15 SNPs significantly associated with five fruit-related traits. Alternatively, seven significant SNPs were detected for only two traits using diploid genotypes, with two SNPs overlapping with the tetraploid scenario. Our results showed that the importance of tetraploid models varied by trait and that the use of diploid models has hindered the detection of SNP-trait associations and, consequently, the genetic architecture of some commercially important traits in autotetraploid species. Furthermore, 14 SNPs co-localized with candidate genes, five of which lead to non-synonymous amino acid changes. The potential functional significance of these SNPs is discussed.

Keywords: autopolyploid, allelic dosage, SNP calling, genetic association, gene action, breeding, *Vaccinium*

INTRODUCTION

Polyploidy is a widespread phenomenon among the flowering plants. Rounds of ancient and recent polyploidization events have been shaping the genomes and the evolutionary trajectories of plant lineages, driving phenotypic diversification (Adams and Wendel, 2005; Paterson, 2005; Jiao et al., 2011; Blischak et al., 2016). Expansion of phenotypic range and novel phenotypes often arise with polyploidization (Spoelhof et al., 2017). The genomic redundancy created by polyploidy allows relaxed selective constraints and functional divergence of gene copies, which can generate new phenotypes in the long-term evolutionary process (Adams and Wendel, 2005; Comai, 2005). Immediate phenotypic effects of polyploidy are also observed compared to their diploid progenitors, such as increased cell and organ size, changes in flowering time, and greater vigor and biomass (Osborn et al., 2003; Tamayo-Ordóñez et al., 2016). The molecular mechanisms contributing to phenotypic variation shortly after polyploidization are not well-understood, but probably involve more complex genetic and epigenetic effects of higher allelic dosage and heterosis (Osborn et al., 2003; Jackson and Chen, 2010; Renny-Byfield and Wendel, 2014; Fort et al., 2016). For example, genome-wide gene expression studies in resynthesized polyploid plants and yeasts have shown ploidy-dependent gene expression alterations, which likely affect the phenotype (Guo et al., 1996; Galitski et al., 1999; Osborn et al., 2003; Pumphrey et al., 2009; Jackson and Chen, 2010).

Polyploids exhibiting new phenotypic traits can outperform their diploid counterparts, occupy new niches, and become ecologically and agriculturally important (Tamayo-Ordóñez et al., 2016; Spoelhof et al., 2017). Many important crops are polyploids with varied ploidy levels and mode of origin (i.e., auto- or allopolyploids). However, despite the economic importance of polyploids and the impact that ploidy can have in the phenotypic expression, the effects of allelic dosage on quantitative traits remain largely unexplored. Most genetic studies in polyploids have so far relied on diploid models to simplify the polyploid data. The complex nature of polyploid genetic data (e.g., multiple alleles and mixed inheritance patterns) has hindered the understanding of genetic architecture of important traits (Dufresne et al., 2014). Moreover, molecular techniques and statistical methodologies were also constraints for polyploids, such as the challenge to define the allelic dosage (Garcia et al., 2013; Lu et al., 2013; Dufresne et al., 2014; Li et al., 2014a; Annicchiarico et al., 2015; Uitdewilligen et al., 2015; Schulz et al., 2016).

Due to the advances in new genotyping technologies, it is now possible to generate high-density single nucleotide polymorphism (SNP) data and evaluate the relative abundance of each allele based on read sequencing depth to infer the allelic dosage. Genome-wide association studies (GWAS) that consider allelic dosage can help uncover the genetic basis of complex traits by considering more realistic genetic models, and hence reducing the signal-to-noise ratio (Garcia et al., 2013; Grandke et al., 2016). Moreover, the effect of the genotype classes on the phenotypic variation can be tested under different gene action models to gain additional insights into additive and non-additive effects (Rosyara et al., 2016). The present study aimed to understand how modeling the allelic dosage influences the identification of SNPs significantly associated with blueberry fruit-related traits through GWAS analyses.

Blueberry has been recognized worldwide for its health benefits, becoming one of the crops with the highest consumer demand and productive trends (USDA, 2016). During blueberry improvement in the United States, interspecific hybridizations have been used for the development of "southern" highbush cultivars adapted to warmer climates. Crosses primarily involved the autotetraploid "northern" highbush blueberry (*Vaccinium corymbosum* L.) and the diploid evergreen blueberry (*V. darrowii* Camp) (Sharpe and Darrow, 1959). Tetraploid hybrids were achieved by the occurrence of unreduced gametes during pollen formation in the diploid species (Ortiz et al., 1992). Despite interspecific hybridizations, blueberry cultivars are considered autotetraploids with non-preferential bivalent chromosome pairing during meiosis and the absence of chromosome structural differentiation (Qu and Hancock, 1995; Qu et al., 1998; Lyrene et al., 2003). The conventional breeding program employs phenotypic recurrent selection, and the release of a new cultivar can take up to 15 years (Hancock et al., 2008). In a perennial polyploid species, such as blueberry, marker-assisted selection has the potential for accelerate the cultivar development process. In this sense, the GWAS analyses can also assist in the identification of causal polymorphisms or molecular markers associated with fruit-related traits relevant for blueberry breeding. The objective of this study is two-fold: (i) to compare the effects of diploid and tetraploid marker calling in population genetics and GWAS analysis; and (ii) to perform the first GWAS analysis for fruit-related traits in southern highbush blueberry.

MATERIALS AND METHODS

Plant Material and Trait Phenotyping

The southern highbush blueberry population used in this study was generated as part of the breeding program at the University of Florida. For this study, 124 controlled crosses were made among 148 selected parents in February 2011. Seeds from each cross were cold-stratified for 5 months and planted in a greenhouse as a family in 2 L pots in November 2011. One hundred seedlings from each family were later transplanted to a high-density nursery (~20,000 plants per 0.2 ha) in a row-column design at the University of Florida Plant Science Research and Education Unit in Citra, Florida. In May 2013, a first round of selection was performed. Unselected plants were removed from the field and the remaining individuals constituted the 1,575 plants from 117 crosses used in this study.

The phenotypic evaluations were conducted during fruit ripening (6 weeks from the beginning of April to mid May 2014) and flowering (January 2015) periods when the plants were in their third growing season. Eight fruit-related traits were measured: weight, size, firmness, stem scar diameter, pH, soluble solids content, flower bud density, and yield. Yield was evaluated using a 1-to-5 rating scale, where 1 indicates none or very few berries on the plant and 5 is a yield comparable to standard commercial cultivars. The flower bud density refers to the number of flower buds on the top 20 cm of one representative upright shoot from the main stem, and was reported as number of buds per 20 cm of shoot. For the fruit traits, the average of five berries randomly selected from each genotype was calculated. Weight (g) was measured using an analytical scale (CP2202S, Sartorius Corp., Bohemia, NY). The same five berries were equatorially oriented to measure fruit size diameter (mm) and firmness (g*mm^{-1} compression force), with a minimum and maximum force threshold of 50 and 350 g, respectively, using the Firm-Tech II (BioWorks Inc., Wamego, KS). The picking stem scar was positioned upward on a tray in a light box with a digital SLR camera (Pentax K-x, Ricoh Imaging, Denver, CO) placed 50 cm above the berry. A ruler was also placed in each image as a size reference. The images were uploaded into FIJI (Schindelin et al., 2012), the scale was set using the ruler, and the scar diameter (mm) was measured for each berry. The blueberry juice was used to measure traits related to sensory quality. The soluble solids content (°Brix), an approximate surrogate measure of sugar content, was assessed using a digital pocket refractometer (Atago U.S.A, Inc., Bellevue, WA). The juice pH was measured using a glass pH electrode (Mettler-Toldeo, Inc., Schwerzenbach, Switzerland).

Capture-Seq Genotyping and SNP Calling

Total genomic DNA was extracted from leaf tissue of each plant using the E-Z 96 PlantDNAKit (Omega Bio-Tek, Norcross, GA). Genotyping was performed by RAPiD Genomics (Gainesville, FL, USA) using sequence capture. Briefly, 31,063 custom-designed biotinylated probes of 120-mer were developed based on the scaffolds of the blueberry draft genome sequence (2013 version) (Bian et al., 2014; Gupta et al., 2015). Sequencing was carried out in the Illumina HiSeq2000 platform using 100 cycle paired-end runs. Raw reads were first trimmed for minimum base quality of 20, demultiplexed, and barcodes were removed. Subsequently, reads were aligned to the blueberry genome (2013 version) using BWA v.0.7.12 (Li and Durbin, 2009).

Polymorphisms and genotypes were called using FreeBayes v.1.0.1, selecting the diploid (-p 2) and the tetraploid (-p 4) options (Garrison and Marth, 2012). Genotypes were represented by the count of alternative alleles. Therefore, for the diploid calling, genotypes were coded as 0 (AA), 1 (AB), or 2 (BB), where "A" and "B" refers to the reference and alternative alleles, respectively. The genotypes for the tetraploid calling were coded as 0 for nulliplex (AAAA), 1 for simplex (AAAB), 2 for duplex (AABB), 3 for triplex (ABBB), and 4 for quadruplex (BBBB). We performed a sample filtering by excluding individuals with more than 90% of missing data across SNPs (sample call rate = 0.9). SNPs were further filtered by: (i) minimum depth of coverage of 40; (ii) minimum genotype quality score of 10; (iii) only biallelic locus; (iv) maximum missing data of 0.7; (v) minor allele frequency of 0.05. The remaining missing genotypes were imputed with the mode of each locus as suggested by Rosyara et al. (2016).

Population Genetics Analyses

Population genetics parameters were computed considering the polyploid and diploid scenarios. We estimated: (i) allele frequency; (ii) heterozygosity; (iii) linkage disequilibrium (LD) decay; and (iv) population structure. The allele frequency for each locus was obtained by counting the number of alternative alleles, divided by sample size, and ploidy level. The observed heterozygosity was calculated as a fraction of the number of heterozygote classes by the total number of loci. Pearson correlation tests (r^2) were performed for pairwise LD estimation within scaffolds. All scaffolds were pooled to plot a genome-wide LD decay and boxplots of r^2 values for categories of marker distances. The decay of LD over genetic distance was determined as the mean distance associated with an empirical LD threshold of $r^2 = 0.2$. To assess the genetic structure of blueberry population, the Principal Components Analysis (PCA) was performed using the marker-based relationship matrix as input. Diploid and tetraploid genomic relationship matrices were computed with the AGHmatrix R-package (Amadeu et al., 2016). The Discriminant Analysis of Principal Components (DAPC) was conducted to cluster genetically similar individuals using the Bayesian Information Criterion (BIC) to select the best supported model, as implemented in the R package adegenet v. 1.3-1 (Jombart and Ahmed, 2011).

GWAS Analyses

The SNP-trait association analyses were based on a linear mixed model, accounting for population structure (**Q**) and relative kinship (**K**) matrices as implemented in the GWASpoly R-package (Rosyara et al., 2016). The **Q**+**K** linear mixed model was:

$$y = ZS\tau \; + \; ZQv \; + \; Zu \; + \; \varepsilon$$

where **y** is a vector of observed phenotypes; ε is a vector of random residual effects, with a multivariate normal distribution with a zero mean vector and an identity variance-covariance (VCOV) matrix; **v** is a vector of sub-populations effects, with incidence matrix **Q**; and **u** is a random polygenic effect, with a multivariate normal distribution with a zero mean vector and VCOV matrix proportional to a kinship matrix (**K**-matrix). The **Z** incidence matrix maps genotypes to observations, and the SNP effects are represented by the τ fixed vector. As pointed out by Rosyara et al. (2016), the matrix **S** depends on the genetic model assumed. In order to compare diploid and tetraploid pipelines, the **Q**+**K** model was implemented in both scenarios. For tetraploid, the **K**-matrix was constructed assuming tetrasomic inheritance (Slater et al., 2013), while for the diploid model it was built considering the algorithm proposed by VanRaden (2008). Both matrices were computed using the AGHmatrix R-package (Amadeu et al., 2016). To correct for population structure, PCA analysis was computed internally using the GWASpoly package and the four principal components were further used in GWAS analyses.

Eight gene action models were tested for the tetraploid genotype calling: general, additive, simplex dominant alternative (simplex-dom-alt), simplex dominant reference (simplex-dom-ref), duplex dominant alternative (duplex-dom-alt), duplex dominant reference (duplex-dom-ref), diplo-additive, and diplo-general. According to Rosyara et al. (2016), the general type of genetic model allows the SNP effect for each genotypic class to be arbitrary and statistically equivalent. In the additive model the SNP effect is proportional to the dosage of the minor allele. In the simplex dominant models, all the heterozygotes (AAAB, AABB, ABBB) are equivalent to one of the homozygotes (AAAA or BBBB). In the duplex dominant models, the duplex state (AABB) has the same effect as either the simplex (AAAB) and nulliplex (AAAA) or the triplex (ABBB) and quadriplex (BBBB) states. In the diploidized models (diplo), all heterozygous classes have the same effect, resembling a traditional diploid dosage model (AA, AB, BB), and have gene action models encompassing the general and additive effects. The diploid genotype calling was also used for GWAS analyses, using the following gene actions: diplo-general, diplo-additive, simplex-dom-alt, and simplex-dom-ref.

Correction for multiple testing using a q-value threshold of 0.05 was applied to determine significant associations using the q-value R-package (Storey and Tibshirani, 2003). We also explored more and less conservative thresholds for declaring significance by using Bonferroni correction of 0.05 and q-value of 0.1, respectively. QQ-plots were used to evaluate the presence of confounding factors leading to an excess of significant associations.

The proportion of phenotypic variation explained by significant SNPs was approximated by the coefficient of determination (R^2). The R^2 was estimated considering a linear regression model that included the first four principal components from PCA analyses, the SNP marker parameterized in accordance with the gene action and a vector of random residual effects.

Candidate Gene Mining

SNPs were characterized *in silico* for their genomic position and functional effect. SNPs were annotated using snpEff v.4.3 (Cingolani et al., 2012), using the blueberry draft genome (2013 version) and gene predictions. Predicted gene models were retrieved from the bitbucket repository https://bitbucket.org/lorainelab/blueberrygenome (Gupta et al., 2015). Candidate genes surrounding significantly associated SNPs were annotated using the Blast2GO tool with BLASTp search against the non-redundant protein database (Götz et al., 2008). We also searched for *Arabidopsis thaliana* v. TAIR10 orthologs using Phytozome v.12.1 BLASTp tool (https://phytozome.jgi.doe.gov).

RESULTS

Phenotypic Variation

A total of 1,575 blueberry plants from 117 crosses were phenotyped for eight fruit-related traits (yield, flower bud density, fruit weight, firmness, size, soluble solids content, pH, and scar diameter). Most traits followed a normal distribution, except yield which was evaluated on a 1-to-5 rating scale, and flower bud density which followed a Poisson distribution (**Figure 1**). High phenotypic correlation was only found between berry size and weight ($r = 0.94$) (**Figure 1**).

Genotypic Data

After filtering the genotypic data, a total of 1,557 individuals and 77,496 SNPs were maintained for the diploid analyses; while 1,559 individuals and 80,591 SNPs were considered for tetraploid analyses. SNPs were sampled throughout the genome, although not evenly distributed, which was expected due to the target design strategy used in this study (**Figure 2**).

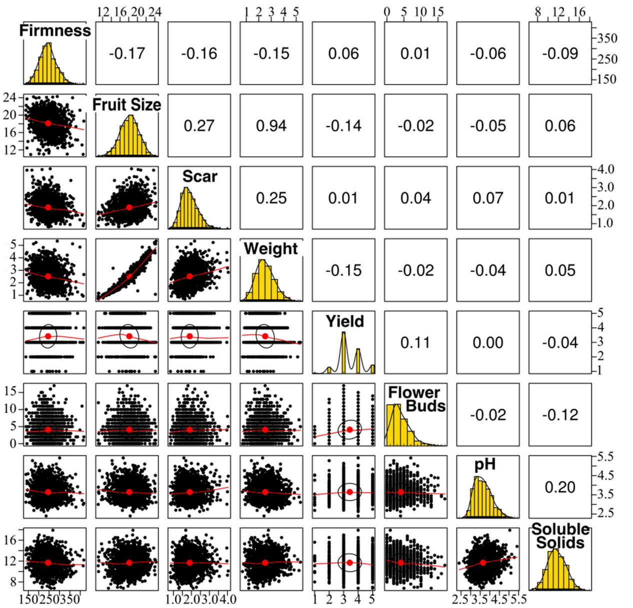


FIGURE 1 | Phenotypic distribution and correlation of eight blueberry fruit-related traits for 1,575 individuals. Plots in the diagonal show the frequency distribution of each trait. Pearson's correlation coefficient between traits are indicated above the diagonal. Scatter plots below the diagonal illustrate the underlying relationship between traits.

Tetraploid and diploid pipelines identified 74,941 common SNPs (around 95% of overlap). We assumed that the differences between pipelines were due to the algorithms implemented in the Freebayes software, which considers different criteria to define a SNP in each parameterization. As a consequence of the high degree of overlap, few differences were observed regarding the position and functional characterization of the SNPs in the blueberry genome (**Figures 3A,B**). Most SNPs were detected in non-coding regions; around 7% targeted exonic regions, mostly causing missense mutations. The distribution of the alternative allele frequency across loci was also similar for both approaches (**Figure 3C**), with the tetraploid model showing the mean allele frequency slightly lower (0.25 vs. 0.27). The main difference between the tetraploid and diploid scenarios was on the genotype calling (**Figure 3D**). For biallelic SNPs in autotetraploids, there are five possible genotypes, with three possible heterozygous states. For diploids, there are only three possible genotypes, with one heterozygous class. The probabilistic assignment of genotypes based on sequence read depth led to a higher heterozygosity for tetraploid compared to diploid genotype calling (0.42 vs. 0.34).

Linkage Disequilibrium and Population Structure

The LD and population structure were consistent between the ploidy models. The trend of LD decay in the blueberry

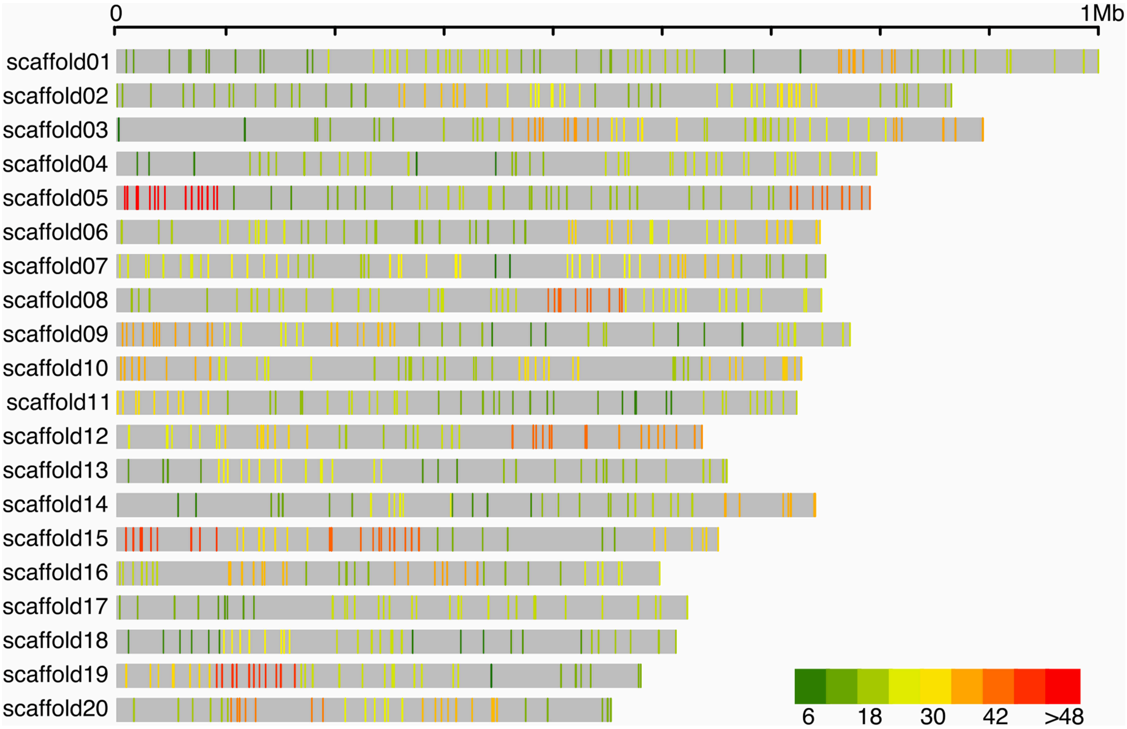


FIGURE 2 | Distribution of filtered SNPs from the tetraploid pipeline in 100 kb windows across the 20 largest blueberry scaffolds (gray). The x-axis represents the distance in base pairs.

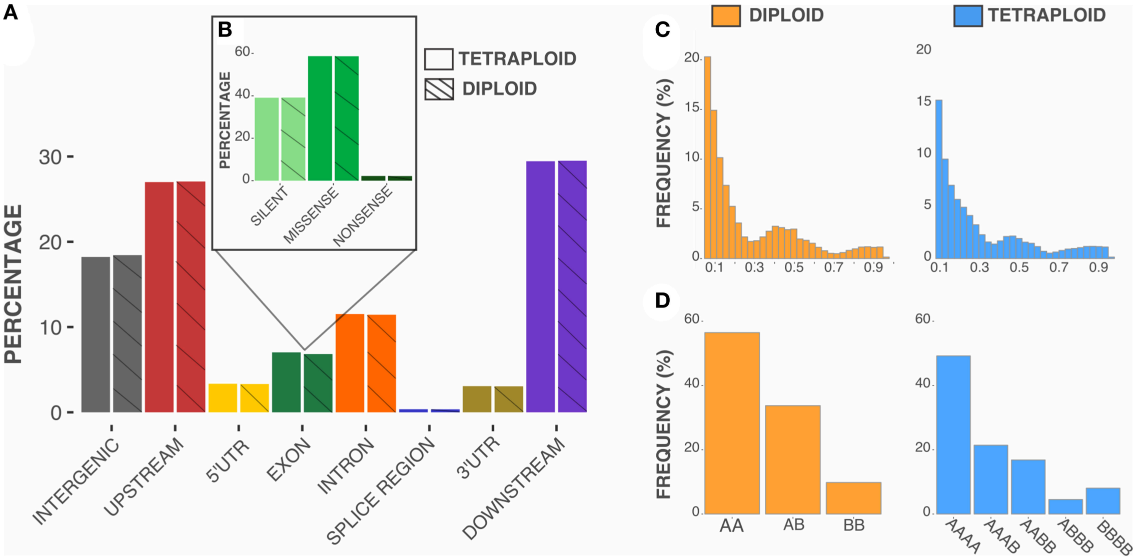


FIGURE 3 | Characterization of SNPs identified in diploid and tetraploid pipelines. **(A)** Percentage of SNPs located in distinct genomic regions. Upstream and downstream regions refer to distances less than 5 kb from surrounding genes. **(B)** Functional effects of SNPs located in exonic regions. **(C)** Distribution of alternative ("B") allele frequency. **(D)** Distribution of genotypic frequencies across loci.

breeding population can be observed by the r^2 distribution across categories of base pair distances between SNPs in **Figure 4**. At the significance threshold ($r^2 = 0.2$), the LD decay presented significant correlation between markers 73 Kb apart for the diploid model and 80 Kb apart for the tetraploid model (Supplementary Figure 1). In order to verify the possible influence of the population structure in the GWAS analysis, we performed PCA and DAPC cluster analyses. The results for both

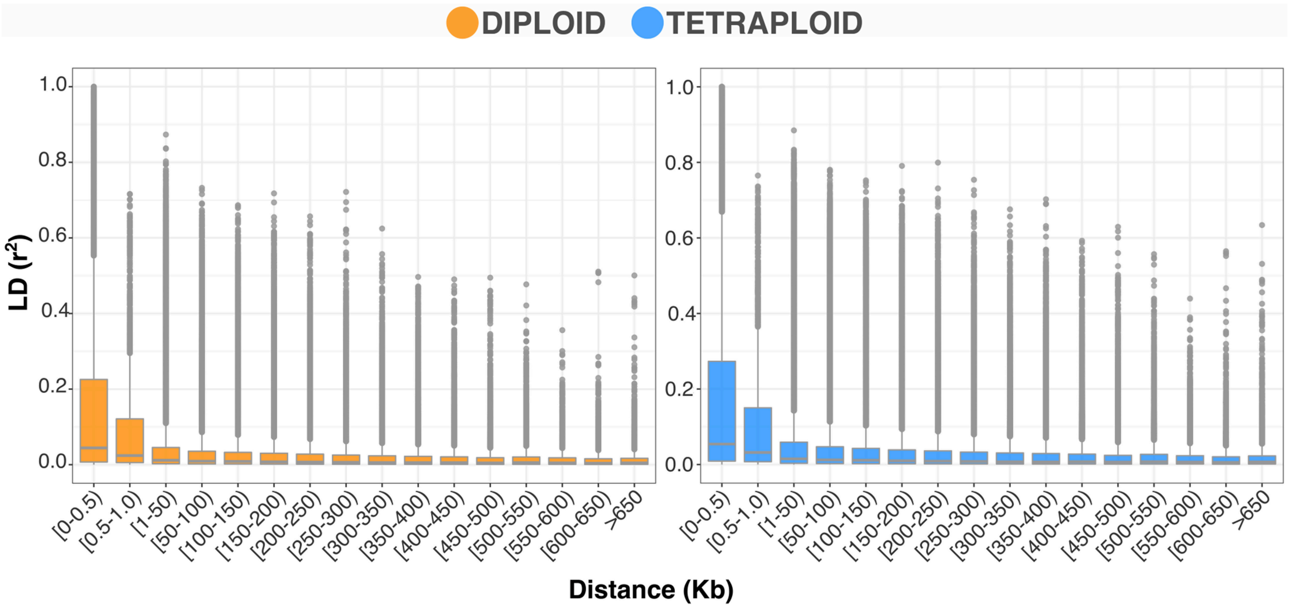


FIGURE 4 | Boxplots showing the trend of LD decay as a relationship between r^2 measures at different intervals of marker distances (Kb) for diploid and tetraploid standardizations.

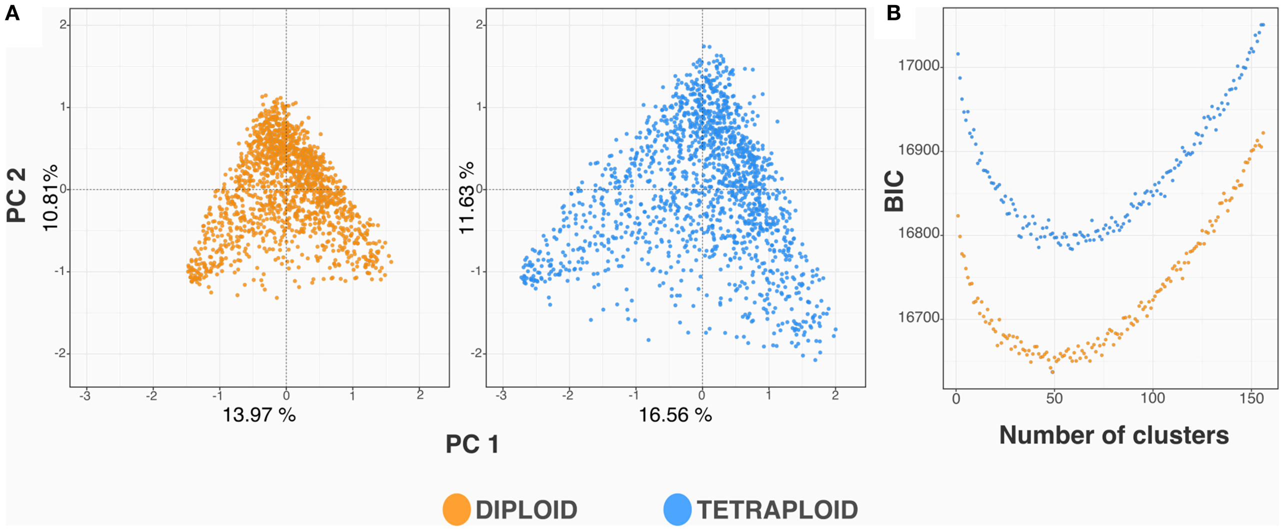


FIGURE 5 | Population structure of a blueberry breeding population of 117 full-sib families performed using diploid and tetraploid pipelines. **(A)** 2D-PCA plots performed using the diploid and tetraploid marker-based relationship matrix. Each individual is represented by a point in the hyperspace defined by the eigenvectors of the first and second principal components. **(B)** Bayesian Information Criterion (BIC) for number of clusters ranging from 0 to 156.

standardizations were very similar, with the tetraploid matrix explaining slightly more of the population genetic variation (28.19 vs. 24.78%) (**Figure 5A**). The comparison of the BIC values for the DAPC analysis suggested the presence of 50 groups in the population (**Figure 5B**), which showed similarities with the pedigree recorded in the population. Hence, in the GWAS analyses, we used the PCA scores to control for population stratification and the genomic relationship matrix to control for cryptic relatedness.

Associations Detected by Polyploid and Diploid Gene Action Models

We performed GWAS analyses for eight fruit-related traits using the **Q+K** linear mixed model. A total of 77,496 and 80,591 SNPs were regressed individually in the diploid and tetraploid GWAS models, respectively. Manhattan plots displaying the significance threshold for each locus in their genomic location are shown in Supplementary Figures 2, 3. The inspection of QQ-plots did not show evidences of systematic bias in any trait or model evaluated (Supplementary Figures 4, 5).

Association analyses using the tetraploid genotypes and a *q*-value threshold of 0.05 allowed the identification of 23 significant SNPs associated with five traits and 11 were also significant after Bonferroni correction (**Table 1**). Six SNPs were identified by more than one gene action model. A total of 15 distinct SNPs were identified: seven for fruit size, two for scar diameter, three for soluble solids content, one for pH, and two for flower bud density (**Figure 6A**, **Table 1**). For fruit size, soluble solids content, and pH traits, dominance models were effective for detecting at least one association. However, the *general* model was the most effective at detecting associations. This class of model assumes that each genotype has its own effect and hence encompasses different gene actions. The inspection of the phenotypic variation across genotypes for significant SNPs identified by the *general* model suggested degrees of overdominance for some traits (e.g., see SNPs *scaffold13749-868* and *scaffold00818-130228* for flower bud density trait) (Supplementary Figure 6). Under a less conservative threshold, the number of distinct associations increased from 15 (*q*-value<0.05) to 37 (*q*-value<0.1) and new associations were detected for fruit weight and firmness traits (**Figure 6A**, Supplementary Table 1). It is

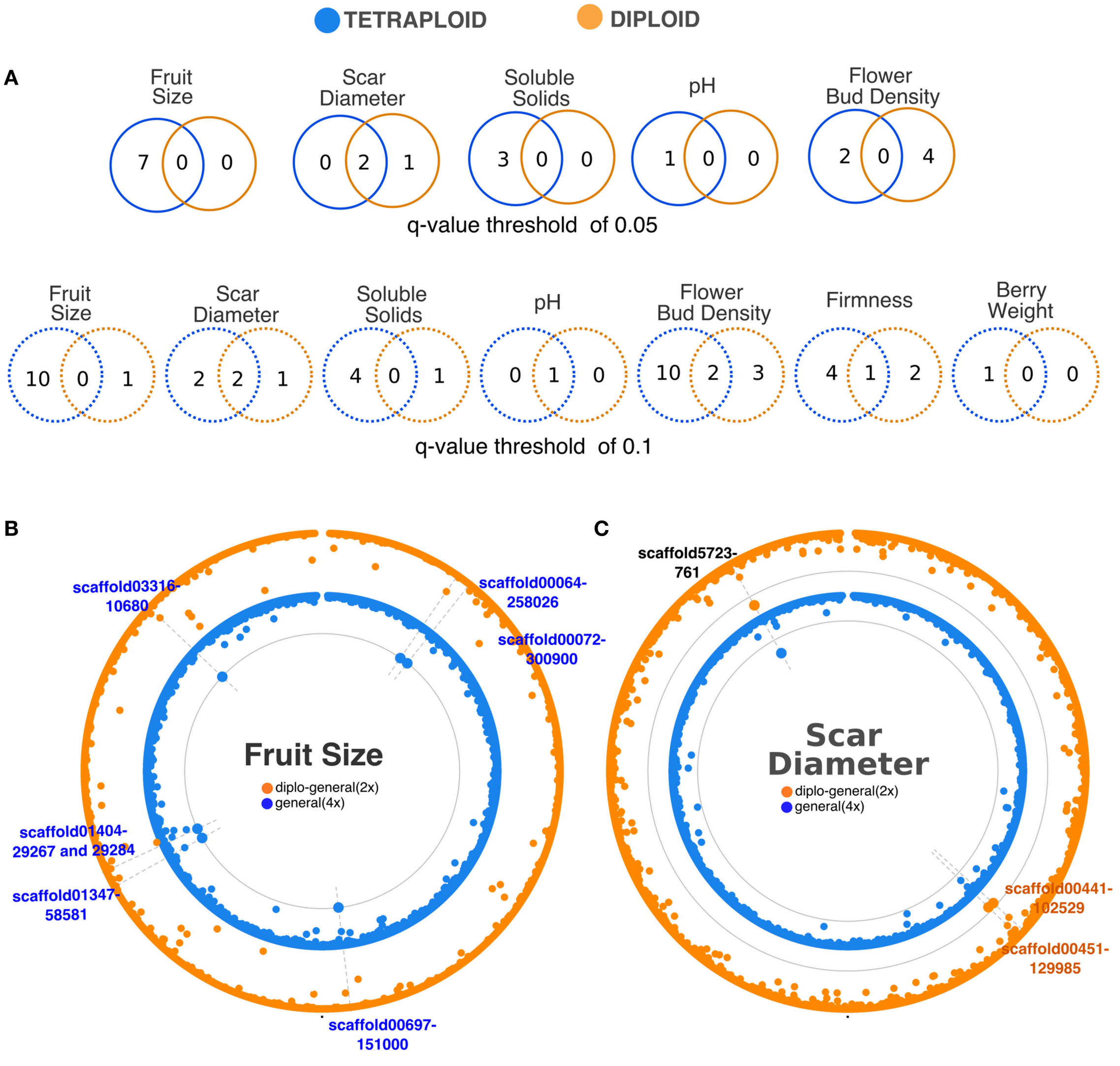


FIGURE 6 | SNP-trait associations detected by modeling tetraploid and diploid genotype callings. **(A)** Venn-diagrams comparing the number of distinct SNPs associated with fruit-related traits in the diploid and tetraploid scenarios under *q*-value thresholds of 0.05 (continuous lines) and 0.1 (dashed lines). **(B)** Circular Manhattan plot for fruit size. Outer and inner layers represent the *diplo-general* and *general* models fitted using diploid (2x) and tetraploid (4x) pipelines, respectively. **(C)** Circular Manhattan plot for scar diameter. Outer and inner layers represent the *diplo-general* and *general* models fitted using the diploid (2x) and tetraploid (4x) pipelines, respectively. SNPs were concatenated by their position in the genomic scaffolds and are displayed along the circular Manhattan plots according to their adjusted *p*-value. The significance threshold (*q*-value = 0.05) is represented by the gray circle in each layer. Vertical dashed gray lines highlight the significant SNPs. The names of significant SNPs are listed outside of the plot. SNPs identified for diploid and tetraploid pipelines are in orange and blue, respectively; while the common SNP identified in both pipelines is in black.

also noteworthy that the same SNP located at scaffold00697, position 151000, was detected as significantly associated with fruit size and fruit weight, the two highly correlated traits (**Figure 1**).

Considering the diploid genotype calling and a q-value threshold of 0.05, we detected seven significant SNPs associated with two fruit-related traits (**Table 1**). Out of these, one association was significant after Bonferroni correction. We found three distinct SNPs associated with scar diameter and four with flower bud density (**Figure 6A**, **Table 1**). The *general* model was the most effective for all traits. Under a less conservative threshold, the number of distinct associations increased from 7 (q-value < 0.05) to 14 (q-value < 0.1) and new associations were detected for berry size, firmness, pH, soluble solids content traits (**Figure 6A**, Supplementary Table 1).

Overall, more SNP-trait associations were identified by modeling the genotypes as tetraploid than as diploid (**Figures 6A,B**). Associations for fruit size, soluble solids content, and pH were only detected using tetraploid models, considering a q-value threshold of 0.05. However, there were four SNPs for flower bud density and one for scar diameter that were only detected by modeling diploid genotypes. Moreover, both models were able to detect the same two SNPs for scar diameter (**Figure 6C**, **Table 1**). No significant association was found for firmness, fruit weight, and yield traits with any ploidy and model tested under this moderate threshold.

Candidate Genes Underlying Fruit-Traits Variation

We identified candidate genes flanking SNPs significantly associated with traits based on the annotation of the blueberry genome (see **Table 1** for q-value < 0.05 and Supplementary Table 1 for q-value < 0.1). Among the protein-coding genes surrounding the seven distinct SNPs associated with fruit size trait, we found a putative lipase (*CUFF.5533.1*), a RING-type E3 ubiquitin ligase (*CUFF.6059.2*), a xyloglucan endotransglucosylase (*CUFF.38641.1*), a hypersensitive-induced response protein 1 (*gene.g14573.t1*), and a chloroplast rhomboid-like protease (*CUFF.39364.1*). Two SNPs in high LD and few base pairs apart were located at the gene encoding the chloroplast RHOMBOID-like protease, one of them leading to a missense mutation (**Figure 7**).

For scar diameter, three distinct SNP-trait associations were detected. Annotation was found only for one of the surrounding genes, which encoded a pentatricopeptide repeat-containing protein (*CUFF.20851.1*).

Three significant SNPs were found for solid soluble content. Two SNPs occurred at genes potentially encoding proteins with a role in the ubiquitin-mediated protein degradation pathway: a ubiquitin-activating enzyme E1 (*CUFF.53548.1*) and an E3 ubiquitin ligase (*CUFF.16799.1*).

For the flower bud density trait, six significant SNPs were found, with four potentially causing missense mutations. Out of those, two SNPs in high LD were located at a gene encoding a zf-RVT domain-containing protein (*CUFF.60704.1*), one at a gene encoding heat shock protein hsp83-90 (*CUFF.13871.1*), and another at a gene encoding a kinase U-box domain-containing protein (*CUFF.57663.1*).

For pH trait, no functional annotation was found for the flanking gene.

DISCUSSION

GWAS analyses in autopolyploids impose additional steps not required in diploids, including the estimation of allele copy number and usage of genetic models that account for dosage effects (Garcia et al., 2013; Dufresne et al., 2014; Rosyara et al., 2016). To circumvent this problem, an alternative has been to use knowledge and methods applied to diploid species in polyploid analyses (Mollinari and Serang, 2015). In this work, we have demonstrated that assuming a diploid parameterization onto a tetraploid species affects the results of a GWAS study. Furthermore, this study is the first to utilize association genetics to understand the genetic architecture and molecular basis of fruit-related traits in blueberry.

How Does Ploidy Affect Population Parameter Estimation?

Prior to performing a SNP-trait association analysis, a detailed understanding of population structure and linkage disequilibrium is essential (Flint-Garcia et al., 2003). Therefore, we compared diploid and polyploid pipelines in terms of marker characterization and estimation of population genetic parameters.

The high degree of overlap between SNP loci identified by both pipelines suggested that the SNP calling step is not drastically affected by ploidy level. However, differences were observed in the genotype calling step, which affected the magnitude of the population parameters. The lower heterozygosity estimated by using diploid (0.34) rather than tetraploid (0.42) genotypes indicates that diploid standardization may cause an underestimation of the heterozygosity rates. Although heterozygosity is a populational parameter and therefore depends on the genetic background under analysis, the heterozygosity estimated in the tetraploid standardization is more in accordance with previous results reported for blueberry (Debnath, 2014; Tailor et al., 2017). Tetraploid highbush blueberry is primarily an outcrossing species with early-acting inbreeding depression (Krebs and Hancock, 1990). Therefore, higher levels of heterozygosity are indeed expected. Moreover, it is reasonable to assume a greater degree of heterozygosity in autopolyploid species in general, since more alleles at one locus are expected when compared to diploids (Gallais, 2003). High levels of heterozygosity have been reported in polyploid species due to its associated benefits, including buffering of deleterious mutations and heterosis (Comai, 2005).

In terms of population-based genomic association studies, it is well-known that population structure is one factor that can result in spurious associations, i.e., associations between a phenotype and markers that are not linked to any causative

TABLE 1 | List of SNP-trait associations detected by diploid and tetraploid models under *q*-value of 0.05 and the candidate genes flaking significant SNPs.

Ploidy	Trait	Gene action model	Scaffold	Position	Ref	Alt	Phen Var (%)	Gene	BLASTp description	BLASTp Arabidopsis TAIR10	Variant type
TETRAPLOID	Fruit Size	General*, diplo-general	scaffold00064	258026	T	C	3.74 | 3.50	CUFF.5533.1	Lipase, alpha/beta-hydrolase	AT1G73920.1	5′UTR premature start codon
	Fruit Size	General	scaffold00072	300900	A	C	3.60	CUFF.6059.2	RING-type E3 ubiquitin ligase	AT5G08750.2	Intron variant
	Fruit Size	General	scaffold00697	151000	C	T	3.44	gene.g14573.t1	Hypersensitive-induced response protein 1 -SPFH/Band 7/PHB domain	AT5G62740.1	Intron variant
	Fruit Size	General	scaffold01347	58581	C	G	3.79	CUFF.38641.1	xyloglucan endotransglucosylase/ hydrolase protein 23	AT4G30270.1	Silent - Ser14Ser
	Fruit Size	Simplex-dom-ref*, general	scaffold01404	29267	C	G	4.21 | 4.57	CUFF.39364.1	Chloroplastic RHOMBOID-like protein 10	AT1G25290.2	Missense - Gln89Glu
	Fruit Size	Simplex-dom-ref*, general	scaffold01404	29284	T	C	4.15 | 4.60	CUFF.39364.1	Chloroplastic RHOMBOID-like protein 10	AT1G25290.2	Silent - Ala94Ala
	Fruit Size	General	scaffold03316	10680	G	A	3.58	NA	NA		Intergenic
	Scar Diameter	Diplo-general	scaffold00451	129985	T	C	3.18	CUFF.21165.1	NA	NA	Silent - Gly5Gly
	Scar Diameter	General*, diplo-general*	scaffold05723	761	A	T	4.02 | 3.51	CUFF.54762.1	NA	NA	5′ UTR variant
	Soluble Solids	Simplex-dom-ref*, diplo-general	scaffold00310	82023	C	T	2.96 | 3.13	CUFF.16799.1	RING/U-box type E3 ubiquitin ligase	AT5G01980.1	Silent - His435His
	Soluble Solids	Simplex-dom-alt*, diplo-additive*, diplo-general*, general	scaffold04772	1107	C	T	2.76 | 2.76 | 2.76 | 3.07	CUFF.53548.1	Ubiquitin-activating enzyme E1	AT2G30110.1	5′ UTR variant
	Soluble Solids	Diplo-general	scaffold06050	414	A	G	2.52	NA	NA	NA	Intergenic region
	pH	Simplex-dom-ref*	scaffold00258	161788	A	G	3.42	CUFF.14779.1	NA	NA	Silent - Gly20Gly
	Flower Buds	General*	scaffold12980	1175	T	G	2.60	CUFF.60704.1	zf-RVT domain-containing protein	AT3G24255.1	Missense - Ile249Leu
	Flower Buds	General	scaffold12980	1177	C	G	1.80	CUFF.60704.1	zf-RVT domain-containing protein	AT3G24255.1	Missense - Ser248Thr
DIPLOID	Scar Diameter	Diplo-general	scaffold00441	102529	A	G	3.32	CUFF.20851.1	Pentatricopeptide repeat (PPR) containing protein	AT5G46100.1	Downstream gene variant
	Scar Diameter	Diplo-general	scaffold00451	129985	T	C	3.24	CUFF.21165.1	NA	NA	Silent - Gly5Gly
	Scar Diameter	Diplo-general*	scaffold05723	761	A	T	3.32	CUFF.54762.1	NA	NA	5′ UTR variant

(Continued)

TABLE 1 | Continued

Ploidy	Trait	Gene action model	Scaffold	Position	Ref	Alt	Phen Var (%)	Gene	BLASTp description	BLASTp Arabidopsis TAIR10	Variant type
	Flower Buds	Diplo-general	scaffold00236	47493	A	G	3.05	CUFF.13871.1	Heat shock protein (hsp83-90)	AT5G52640.1	Silent- Ile31Ile
	Flower Buds	Diplo-general	scaffold00236	47494	A	G	3.03	CUFF.13871.1	Heat shock protein (hsp83-90)	AT5G52640.1	Missense - Ile31Thr
	Flower Buds	diplo-General	scaffold01219	10424	C	T	2.84	54064_g.1	NA		upstream gene variant
	Flower Buds	Diplo-general	scaffold08724	1370	T	A	1.91	CUFF.57663.1	Protein kinase protein, U-box 52 domain-containing	AT1G16760.1	Missense - Asp98Val

*SNPs identified by both diploid and tetraploid models are highlighted in bold. Scaffolds and positions refer to the blueberry genome assembly version 2013. Ref, Reference allele; Alt, Alternative allele; Phen Var, Phenotypic variation explained by the marker; NA, no annotation; *Significant after Bonferroni correction (p < 0.05).*

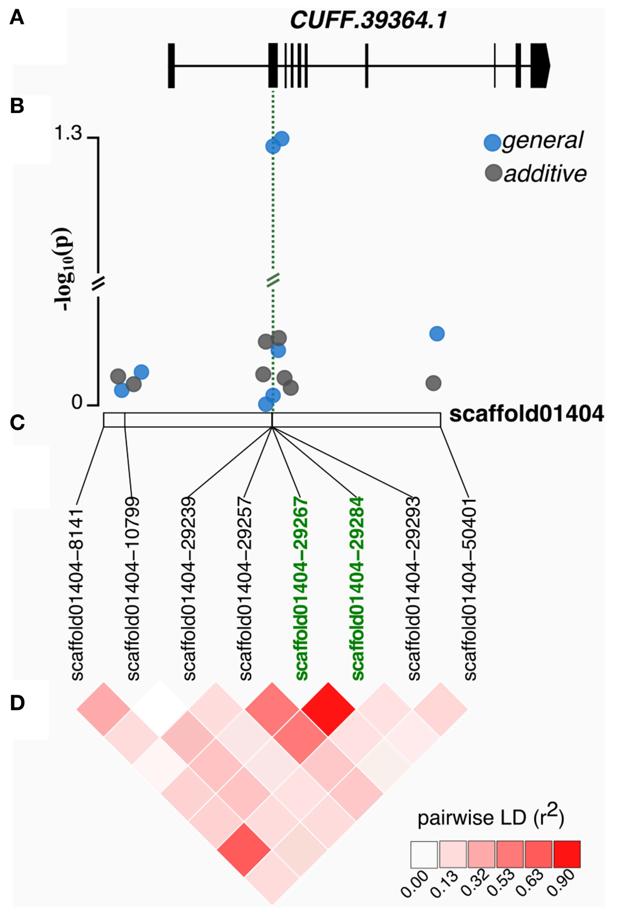


FIGURE 7 | SNP effect on fruit size. **(A)** Candidate gene encoding a chloroplast rhomboid-like protease (*CUFF.39364.1*) where a missense variant was detected in the second exon. **(B)** Significance of SNPs detected by *additive* (gray dots) and *general* (blue dots) gene action models along the scaffold01404. The green dashed line indicates the genic region affected by the two significant SNPs. Double bars indicate out of scale. **(C)** Scaffold positions of the SNPs, highlighting the SNPs associated with the trait in green. **(D)** Pairwise linkage disequilibrium (correlation coefficient r^2) between markers along the scaffold.

loci (Pritchard et al., 2000; Sillanpää, 2011). For diploid and tetraploid pipelines, the most likely number of groups in DAPC analyses were in accordance with the pedigree recorded in the population. Based on the QQ-plot results, we inferred that the first four principal components and the genomic relationship matrices in each parameterization were sufficient to account for sample structure confounders. However, it is noteworthy that this conclusion is limited to our breeding population. In more complex pedigrees, for example, the usage of relationship matrix for autotetraploids might impact the final results (Kerr et al., 2012; Amadeu et al., 2016).

LD is another population parameter that significantly affects GWAS results. Assuming that association analyses rely on non-random association between SNPs and causative genes, determining the extent of LD is important to define strategies in GWAS analyses. For both pipelines, we observed a rapid

LD decay across the blueberry scaffolds. Accordingly, low LD is reported in other outcrossing species (Gupta et al., 2005). For practical purposes, short LD blocks require a higher number of individuals with records and higher marker density in order to identify causal variants (Goddard et al., 2016). Hence, the usage of a high number of individuals and a high throughput genotyping method was consistent with our research scenario. The LD pattern can also provide information about the genetic diversity in our breeding population. Assuming that the expectation of r^2 can be expressed as a function of the effective population size (Ne), faster LD decay is expected as long as Ne increases (Flint-Garcia et al., 2003). Empirically, a short-range LD observed in our population suggests a large Ne value. This is in accordance with the breeding strategy at the University of Florida as parental selection has been performed in order to decrease the inbreeding depression, therefore maintaining genetic diversity (Cellon et al., 2018).

SNP-Trait Associations in Autotetraploid Blueberries

Polyploid studies considering the relative abundance of each allele at a particular locus in the genome allow the testing of more realistic genetic models. For example, the usage of allele dosage has impacted the construction of genetic linkage maps (Mollinari and Serang, 2015), the computation of observed and expected allele frequencies (Dufresne et al., 2014), and the inference of population structure and patterns of historical demography (Blischak et al., 2016). On the bases of genome-wide association studies, our results supported the importance of including allelic dosage to identify significant SNP-trait associations. By modeling tetraploid genotypes under a q-value threshold of 0.05, at least one SNP-trait association was detected for five traits in a blueberry breeding population, and no associations were detected for fruit size, pH, and soluble solids traits when the dosage effect was omitted.

In addition to the allelic dosage, we also tested different gene action models. It is noteworthy that the genotypic value of an individual is estimated differently in polyploid and diploid species. In autotetraploids, the higher number of alleles per locus reflects on different coefficients of dominance, increasing the range of genetic models to describe one-locus genotypic value (Gallais, 2003). In this study, dominance gene actions were addressed on the *simplex* and *duplex dominance* models. *Simplex dominance* represents the first order interaction among alleles and may be modeled regardless of the ploidy. Nevertheless, *duplex dominance* arises when heterozygotes are affected only if they have two unfavorable alleles; therefore, it is a model that can only be tested in polyploid systems. *Duplex dominance* interaction models were detected for associations under q-value threshold of 0.1 for flower bud density and firmness traits. Hence, our results reinforce the importance of considering an autotetraploid parameterization in blueberry.

We also tested "diploidized models" or "pseudodiploid models" using the tetraploid genotype calling, as they are widely-used in polyploid analyses due to straightforward implementation in diploid software (Li et al., 2014b; Biazzi et al., 2017). This parameterization disregards the allele dosage and all heterozygotes are grouped into the same genotypic class, which is at the midpoint between the two homozygotes (Rosyara et al., 2016; Slater et al., 2016). In diploid species, this is equivalent to the additive model (parameterized as {0,1,2} and assuming that the SNP effect is proportional to the dosage of the minor allele). In autotetraploids, this parameterization might be interpreted as a *partial dominance* model suggesting that any order of interaction between alleles reduces the genotypic value (Gallais, 2003; Slater et al., 2016). Our results showed that "diploidized models" were valid for scar diameter, fruit size, and soluble solids traits under a q-value threshold of 0.05. Interestingly, the standard assumption of additivity was not the most appropriate to describe the phenotypic variation observed in blueberry. Divergent results were described in autopolyploid potatoes, for which most of the QTLs were identified considering additive models (Rosyara et al., 2016). Based on our results, we might infer that non-additive effects have a key role in understanding the genetic architecture of blueberry fruit traits.

Although we did not have explicitly approached models addressing partial interactions among alleles, they are potential models to be further implemented in GWAS analyses. *Overdominance* is particularly more complex, since it can be explored by restricting interactions among alleles to different orders (Gallais, 2003). In this study, these genetic assumptions were implicitly considered in the *general* model. *General* model is a generic class that also encompasses other models with no genetic assumptions (Rosyara et al., 2016). Not surprisingly, this model was able to identify the highest number of significant trait-associations, with some overlap with the competing models.

However, considering a q-value threshold of 0.1, significant associations were identified by *simplex* and *duplex* models for soluble solids, flower bud density, and pH, but not by the *general* model. According to Rosyara et al. (2016), there is a trade-off between flexibility and power, because the *general* model requires a higher number of degrees of freedom, resulting in a lower statistical power.

The heritability estimate provided some insights into the results. Heritability is a population parameter that measures the degree of variation in a phenotypic trait that is due to genetic variation (Falconer and Mackay, 1996). Therefore, it is reasonable to expect a positive relation between heritability and ability to detect associations. In the current population, low to mid narrow-sense heritability was found for the traits, varying from 0.16 for flower bud density to 0.57 for scar diameter (for details, see Cellon et al., 2018). In line with this, individual markers explained a small portion of the phenotypic variation (less than 5%). These results suggest that all fruit-related traits analyzed herein are quantitative, which means that phenotypic variation depends on the cumulative actions of many genes with small effects and their interaction with environment.

Biological Insights Into the Genetic Basis of Fruit-Related Traits in Blueberry

Among the significant SNPs associated with blueberry fruit-related traits, some did not lie in protein-coding regions and others caused synonymous changes. In the majority of the GWAS studies in plants, significant associations were also detected for variants in introns, untranslated, or intergenic regions (Ingvarsson and Street, 2011). Many of these variants can be in LD with an untyped causal non-synonymous mutation or might cause changes in gene expression (Gilad et al., 2008). In the case of blueberry, the absence of a high-quality reference genome is an additional challenge for GWAS analysis and biological interpretation. The current available genome is very fragmented and many predicted genes are incomplete (Gupta et al., 2015). Hence, the biological significance of the associations found herein is still limited and speculative, but we point out some insights into the potential molecular mechanisms underlying the variation of each trait.

Larger fruits are a consumer-desired trait in the fresh blueberry market. Among the significant SNPs associated with fruit size, one caused a non-synonymous mutation in the putative gene encoding a chloroplast-located rhomboid-like protease. In *A. thaliana,* the lack of a rhomboid protease was associated with reduced fertility and aberrations in flower morphology (Knopf et al., 2012; Thompson et al., 2012). Changes in floral morphology and development can affect the fruit size and shape, as reported in tomato (Tanksley, 2004). However, to our knowledge, no study has reported the role of a rhomboid-like protease in fruit size variation. Another SNP associated with berry size occurred at a gene encoding a RING-type E3 ubiquitin ligase. Interestingly, a QTL for rice grain width and weight was also mapped in RING-type protein with E3 ubiquitin ligase activity (Song et al., 2007). Song et al. (2007) suggested that this protein negatively regulates cell division by targeting its substrate(s) to proteasome degradation, since its loss of function resulted in increased cell number and larger (wider) rice spikelet hull. Another interesting SNP was the one located in a gene encoding a xyloglucan endotransglucosylase. This enzyme catalyzes the molecular grafting between xyloglucan molecules in the plant cell-wall matrix, allowing expansive cell growth by restructuring the cell wall (Miedes et al., 2011; Ohba et al., 2011). In transgenic tomatoes with modified expression of a xyloglucan endotransglucosylase gene, fruit size was positively correlated with the expression level of this enzyme (Ohba et al., 2011).

The picking scar size also affects blueberry commercialization, as bigger scars increase perishability and pathogen penetration (Parra et al., 2007). Among the associations detected for scar diameter, the most interesting was the SNP detected under a *q*-value of 0.1, upstream of an auxin transporter 3, which controls cellular auxin influx. The major form of auxin IAA (Indole-3-acetic acid) is known to delay fruit abscission from the receptacle by reducing the sensitivity of cells in the abscission zone to ethylene (Blanusa et al., 2005; Kühn et al., 2016). The inhibition of polar auxin transport in grapevine fruitlets resulted in fruit drop (Kühn et al., 2016).

Soluble solid content and pH are important sensory quality factors affecting blueberry fruit flavor. Sweetness perception of fruits depends on the balance between sugars and acids (Cirilli et al., 2016; Farneti et al., 2017). For the sugar content, measured as the soluble solids content, two significant SNPs occurred at genes encoding proteins with a role in the ubiquitin-mediated protein degradation pathway. The attachment of ubiquitin molecules to selected proteins can have diverse regulatory functions, influencing the protein activity, abundance, trafficking, or localization (Stone, 2014). The ubiquitin-proteasomal degradation machinery is also involved in the regulation of sugar signaling pathways, which primarily targets the source-to-sink carbon partitioning (Rolland et al., 2006). The role of proteolysis in controlling sugar accumulation was also reported in tomato fruits (Ariizumi et al., 2011). For pH variation, no annotation was found for the predicted gene harboring the significant SNP, hindering biological insights at this point.

Flower bud density can be useful to estimate potential yield in the next harvest (Salvo et al., 2012). Among the significant associations with this trait, we found SNPs leading to missense mutations. One missense mutation occurred at the gene encoding for a heat shock protein (*hsp83-90*). In *Ipomoea nil* (formely *Pharbitis nil*, the Japanese morning glory), *hsp83* was upregulated upon exposure to a photoperiod that induces flowering (Felsheim and Das, 1992). The heat shock protein Hsp90 was also reported to act as an environmental signal sensor regulating flowering time (Sangster et al., 2007) and flower development (Margaritopoulou et al., 2016). Another missense variant was found at a gene encoding for a protein kinase U-box domain-containing. The U-box domain has a ubiquitin ligase activity and the kinase motif suggests that this protein participates in signal transduction cascades via phosphorylation. The potential ortholog in *Arabidopsis thaliana* (*At1g16760*) is expressed during the pollen stage (Wang et al., 2008).

Fruit firmness is a trait of commercial importance as it directly affects fruit quality, shelf life, and transportability (MacLean and NeSmith, 2011); therefore, it is a key target for blueberry breeding. In this work, we identified associations only when we used a less stringent *q*-value threshold of 0.1; two missense variants were detected. One of the SNPs causing missense mutations was located at a putative ubiquitin-like-specific cysteine proteinase. Recent studies have shown the role of proteolysis in the regulation of fruit ripening in tomato (Wang et al., 2014, 2017). Particularly, a vacuolar cysteine proteinase (*SlVPE3*) was shown to affect the accumulation of numerous ripening-related proteins, acting as a post-transcriptional regulator (Wang et al., 2017). Moreover, Salentijn et al. (2003) found cysteine proteinases differentially expressed between firm and soft strawberry cultivars. The other missense variant associated with firmness was located in a SAM-MTase. SAM-MTases are ubiquitous enzymes that catalyze the transfer of methyl groups from S-adenosyl methionine (SAM) to a myriad of compounds (e.g., DNA, RNA, proteins, sterols, pectin, lignin, flavonoids, phenylpropanoids, and alkaloids) and also act in the biosynthesis pathway of ethylene and polyamines. Many of those compounds have an important role in fruit ripening (Moffatt and Weretilnyk, 2001; Roje, 2006; Teyssier et al., 2008; Singh et al., 2010; MacLean and NeSmith, 2011;

Paul et al., 2012; Van de Poel et al., 2013; Zhang et al., 2015).

Current Challenges and Perspectives of GWAS in the Blueberry Breeding Program

Two of the major challenges faced in this study were the absence of a high-quality genome assembly for blueberry and the allelic dosage calling. We expect that the improvement of genome contiguity might impact the reads alignment quality, providing a more accurate SNP calling and a more precise location of the markers associated with traits. Dosage calling has also been recognized to be a major challenge in genomic studies of polyploid species (Bourke et al., 2018), and it is an area that when fully developed could contribute significantly to association studies in autopolyploids. Population structure is another issue that could be affecting the current results. Controlling for population structure is a standard procedure in GWAS analyses, as we did by using the Q+K model; however, it reduces the statistical power to detect associations when phenotypes strongly correlate with relatedness (Reif et al., 2010; Brachi et al., 2011; Würschum et al., 2012; Ogut et al., 2015; Han et al., 2016; Klasen et al., 2016).

Our results suggested that blueberry fruit quality traits have a complex genetic basis. Therefore, the traditional implementation of marker-assisted selection using our GWAS results seems limited at this point. However, we emphasize that new associations with higher effects could be detected in future GWAS analyses using a complete genome assembly, higher marker density, and more accurate dosage calling method. Alternatively, genomic selection is a promising approach for prediction of complex traits and it is an opportunity for future studies.

CONCLUSION

Altogether, in this study we demonstrated that simplifying tetraploid data as a diploid can have significant consequences in some population genetic parameters and in the ability to detect marker-trait associations. The absence of associations detected by the conventional additive gene action model suggests that non-additive effects might play a key role in understanding the genetic architecture of blueberry fruit traits. Some of the significant SNPs were detected within and around biologically plausible candidate genes. The encoded proteins may act on pathways that affect the traits as suggested by studies in other plant species. However, better gene prediction and functional validation of these genes will further improve our understanding of the variation of fruit-related traits in blueberry.

DATA AVAILABILITY

Phenotypic and genotypic datasets used for diploid and tetraploid analyses are available from the Dyrad Digital Repository (accession number doi: 10.5061/dryad.kd4jq6h).

AUTHOR CONTRIBUTIONS

PM and JO designed the study. CC and JO conducted the field experiment and collected the phenotypic data. CC performed the DNA extraction. MR performed the SNP calling and filtering. LF, JB, and IdB performed the data analyses and interpretation. JB, LF, IdB, and PM wrote the paper. MR and MK provided analytical expertise and edited the manuscript. PM supervised the whole study. All authors read and approved the final version of the manuscript for publication.

FUNDING

This work was funded by the UF royalty fund generated by the licensing of blueberry cultivars. CC was partially supported, while at the University of Florida, by the National Institute of Food and Agriculture, United States Department of Agriculture, under award number 2014-67013-22418 to PM and JO. IdB was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) [PSDE scholarship: 88881.131685/2016-01].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2018.00107/full#supplementary-material

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A, Resende, M. F., and Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. *Plant Genome* 9. doi: 10.3835/plantgenome2016.01.0009

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16:1020. doi: 10.1186/s12864-015-2212-y

Ariizumi, T., Higuchi, K., Arakaki, S., Sano, T., Asamizu, E., and Ezura, H. (2011). Genetic suppression analysis in novel vacuolar processing enzymes reveals their roles in controlling sugar accumulation in tomato fruits. *J. Exp. Bot.* 62, 2773–2786. doi: 10.1093/jxb/erq451

Bian, Y., Ballington, J., Raja, A., Brouwer, C., Reid, R., Burke, M., et al. (2014). Patterns of simple sequence repeats in cultivated blueberries (*Vaccinium* section Cyanococcus spp.) and their use in revealing genetic diversity and population structure. *Mol. Breed.* 34, 675–689. doi: 10.1007/s11032-014-0066-7

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., et al. (2017). Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PLoS ONE* 12:e0169234. doi: 10.1371/journal.pone.0169234

Blanusa, T., Else, M. A., Atkinson, C. J., and Davies, W. J. (2005). The regulation of sweet cherry fruit abscission by polar auxin transport. *Plant Growth Regul.* 45, 189–198. doi: 10.1007/s10725-005-3568-9

Blischak, P. D., Kubatko, L. S., and Wolfe, A. D. (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol. Ecol. Resour.* 16, 742–754. doi: 10.1111/1755-0998.12493

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* 9:513. doi: 10.3389/fpls.2018.00513

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12:232. doi: 10.1186/gb-2011-12-10-232

Cellon, C., Amadeu, R. R., Olmstead, J. W., Mattia, M. R., Ferrão, L. F. V., and Munoz, P. R. (2018). Estimation of genetic parameters and prediction of breeding values in an autotetraploid blueberry breeding population with extensive pedigree data. *Euphytica* 214:87. doi: 10.1007/s10681-018-2165-8

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Cirilli, M., Bassi, D., and Ciacciulli, A. (2016). Sugars in peach fruit: a breeding perspective. *Hortic. Res.* 3:15067. doi: 10.1038/hortres.2015.67

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Debnath, S. C. (2014). Structured diversity using EST-PCR and EST-SSR markers in a set of wild blueberry clones and cultivars. *Biochem. Syst. Ecol.* 54, 337–347. doi: 10.1016/j.bse.2014.03.018

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Falconer, D. S., and Mackay, T. F. C. (1996). *Quantitative Genetics*. New York, NY: Longman Scientific and Technical.

Farneti, B., Khomenko, I., Grisenti, M., Ajelli, M., Betta, E., Algarra, A. A., et al. (2017). Exploring blueberry aroma complexity by chromatographic and direct-injection spectrometric techniques. *Front. Plant Sci.* 8:617.doi: 10.3389/fpls.2017.00617

Felsheim, R. F., and Das, A. (1992). Structure and expression of a heat-shock protein 83 gene of *Pharbitis nil*. *Plant Physiol.* 100, 1764–1771. doi: 10.1104/pp.100.4.1764

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Fort, A., Ryder, P., McKeown, P. C., Wijnen, C., Aarts, M. G., Sulpice, R., et al. (2016). Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in *Arabidopsis thaliana*. *New Phytol.* 209, 590–599. doi: 10.1111/nph.13650

Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999). Ploidy regulation of gene expression. *Science* 285, 251–254. doi: 10.1126/science.285.5425.251

Gallais, A. (2003). *Quantitative Genetics and Breeding Methods in Autopolyploid Plants*. Paris: INRA Editions.

Garcia, A. A., Mollinari, M., Marconi, T. G., Serang, O. R., Silva, R. R., Vieira, M. L., et al. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* 3:3399. doi: 10.1038/srep03399

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint]. arXiv1207.3907.

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24, 408–415. doi: 10.1016/j.tig.2008.06.001

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., and Hayes, B. J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. R. Soc. B*. 283:20160569. doi: 10.1098/rspb.2016.0569

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Grandke, F., Singh, P., Heuven, H. C., De Haan, J. R., and Metzler, D. (2016). Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum. *BMC Genomics* 17:672. doi: 10.1186/s12864-016-2926-5

Guo, M., Davis, D., and Birchler, J. A. (1996). Dosage effects on gene expression in a maize ploidy series. *Genetics* 142, 1349–1355.

Gupta, P. K., Rustgi, S., and Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* 57, 461–485. doi: 10.1007/s11103-005-0257-z

Gupta, V., Estrada, A. D., Blakley, I., Reid, R., Patel, K., Meyer, M. D., et al. (2015). RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. *Gigascience* 4, 1–22. doi: 10.1186/s13742-015-0046-9

Han, S., Utz, H. F., Liu, W., Schrag, T. A., Stange, M., Würschum, T., et al. (2016). Choice of models for QTL mapping with multiple families and design of the training set for prediction of *Fusarium* resistance traits in maize. *Theor. Appl. Genet.* 129, 431–444. doi: 10.1007/s00122-015-2637-3

Hancock, J. F., Lyrene, P., Finn, C. E., Vorsa, N., and Lobos, G. A. (2008). "Blueberries and cranberries," in *Temperate Fruit Crop Breeding*, ed J. F. Hancock (Dordrecht: Springer), 115–150.

Ingvarsson, P. K., and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x

Jackson, S., and Chen, Z. J. (2010). Genomic and expression plasticity of polyploidy. *Curr. Opin. Plant Biol.* 13, 153–159. doi: 10.1016/j.pbi.2009.11.004

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Kerr, R. J., Li, L., Tier, B., Dutkowski, G. W., and McRae, T. A. (2012). Use of the numerator relationship matrix in genetic analysis of autopolyploid species. *Theor. Appl. Genet.* 124, 1271–1282. doi: 10.1007/s00122-012-1785-y

Klasen, J. R., Barbez, E., Meier, L., Meinshausen, N., Bühlmann, P., Koornneef, M., et al. (2016). A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nat. Commun.* 7:13299. doi: 10.1038/ncomms13299

Knopf, R. R., Feder, A., Mayer, K., Lin, A., Rozenberg, M., Schaller, A., et al. (2012). Rhomboid proteins in the chloroplast envelope affect the level of allene oxide synthase in *Arabidopsis thaliana*. *Plant J.* 72, 559–571. doi: 10.1111/j.1365-313X.2012.05090.x

Krebs, S. L., and Hancock, J. F. (1990). Early-acting inbreeding depression and reproductive success in the highbush blueberry, *Vaccinium corymbosum* L. *Theor. Appl. Genet.* 79, 825–832. doi: 10.1007/BF00224252

Kühn, N., Serrano, A., Abello, C., Arce, A., Espinoza, C., Gouthu, S., et al. (2016). Regulation of polar auxin transport in grapevine fruitlets (*Vitis vinifera* L.) and the proposed role of auxin homeostasis during fruit abscission. *BMC Plant Biol.* 16:234. doi: 10.1186/s12870-016-0914-1

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, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., et al. (2014a). Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS ONE* 9:e84329. doi: 10.1371/journal.pone.0084329

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014b). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3* 4, 1971–1979. doi: 10.1534/g3.114.012245

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9:e1003215. doi: 10.1371/journal.pgen.1003215

Lyrene, P. M., Vorsa, N., and Ballington, J. R. (2003). Polyploidy and sexual polyploidization in the genus *Vaccinium*. *Euphytica* 133, 27–36. doi: 10.1023/A:1025608408727

MacLean, D. D., and NeSmith, D. S. (2011). Rabbiteye blueberry postharvest fruit quality and stimulation of ethylene production by 1-methylcyclopropene. *Hortscience* 46, 1278–1281.

Margaritopoulou, T., Kryovrysanaki, N., Megkoula, P., Prassinos, C., Samakovli, D., Milioni, D., et al. (2016). HSP90 canonical content organizes a molecular scaffold mechanism to progress flowering. *Plant J.* 87, 174–187. doi: 10.1111/tpj.13191

Miedes, E., Zarra, I., Hoson, T., Herbers, K., Sonnewald, U., and Lorences, E. P. (2011). Xyloglucan endotransglucosylase and cell wall extensibility. *J. Plant Physiol.* 168, 196–203. doi: 10.1016/j.jplph.2010.06.029

Moffatt, B. A., and Weretilnyk, E. A. (2001). Sustaining S-adenosyl-l-methionine-dependent methyltransferase activity in plant cells. *Physiol. Plant* 113, 435–442. doi: 10.1034/j.1399-3054.2001.1130401.x

Mollinari, M., and Serang, O. (2015). "Quantitative SNP genotyping of polyploids with MassARRAY and other platforms," in *Plant Genotyping Methods in Molecular Biology (Methods and Protocols)*, ed J. Batley (New York, NY: Humana Press), 215–241.

Ogut, F., Bian, Y., Bradbury, P. J., and Holland, J. B. (2015). Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity* 114, 552–563. doi: 10.1038/hdy.2014.123

Ohba, T., Takahashi, S., and Asada, K. (2011). Alteration of fruit characteristics in transgenic tomatoes with modified expression of a xyloglucan endotransglucosylase/hydrolase gene. *Plant Biotechnol. J.* 28, 25–32. doi: 10.5511/plantbiotechnology.10.0922a

Ortiz, R., Vorsa, N., Bruederle, L. P., and Laverty, T. (1992). Occurrence of unreduced pollen in diploid blueberry species, *Vaccinium* sect. Cyanococcus. *Theor. Appl. Genet.* 85, 55–60. doi: 10.1007/BF00223844

Osborn, T. C., Pires, J. C., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H. S., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* 19, 141–147. doi: 10.1016/S0168-9525(03)00015-5

Parra, R., Lifante, Z. D., and Valdés, B. (2007). Fruit size and picking scar size in some blueberry commercial cultivars and hybrid plants grown in SW Spain. *Int. J. Food Sci. Technol.* 42, 880–886. doi: 10.1111/j.1365-2621.2006.01299.x

Paterson, A. H. (2005). Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica* 123, 191–196. doi: 10.1007/s10709-003-2742-0

Paul, V., Pandey, R., and Srivastava, G. C. (2012). The fading distinctions between classical patterns of ripening in climacteric and non-climacteric fruit and the ubiquity of ethylene—an overview. *J. Food Sci. Technol.* 49, 1–21. doi: 10.1007/s13197-011-0293-4

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181. doi: 10.1086/302959

Pumphrey, M., Bai, J., Laudencia-Chingcuanco, D., Anderson, O., and Gill, B. S. (2009). Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics* 181, 1147–1157. doi: 10.1534/genetics.108.096941

Qu, L., and Hancock, J. F. (1995). Nature of 2n gamete formation and mode of inheritance in interspecific hybrids of diploid *Vaccinium darrowi* and tetraploid *V. corymbosum. Theor. Appl. Genet.* 91, 1309–1315. doi: 10.1007/BF00220946

Qu, L., Hancock, J., and Whallon, J. (1998). Evolution in an autopolyploid group displaying predominantly bivalent pairing at meiosis: genomic similarity of diploid *Vaccinium darrowi* and autotetraploid *V. corymbosum (Ericaceae). Am. J. Bot.* 85, 698–703. doi: 10.2307/2446540

Reif, J. C., Liu, W., Gowda, M., Maurer, H. P., Möhring, J., Fischer, S., et al. (2010). Genetic basis of agronomically important traits in sugar beet (*Beta vulgaris L.*) investigated with joint linkage association mapping. *Theor. Appl. Genet.* 12, 1489–1499. doi: 10.1007/s00122-010-1405-7

Renny-Byfield, S., and Wendel, J. F. (2014). Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* 101, 1711–1725. doi: 10.3732/ajb.1400119

Roje, S. (2006). S-Adenosyl-L-methionine: beyond the universal methyl group donor. *Phytochemistry* 67, 1686–1698. doi: 10.1016/j.phytochem.2006.04.019

Rolland, F., Baena-Gonzalez, E., and Sheen, J. (2006). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annu. Rev. Plant Biol.* 57, 675–709. doi: 10.1146/annurev.arplant.57.032905.105441

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2015.08.0073

Salentijn, E. M. J., Aharoni, A., Schaart, J. G., Boone, M. J., and Krens, F. A. (2003). Differential gene expression analysis of strawberry cultivars that differ in fruit-firmness. *Physiol. Plant.* 118, 571–578. doi: 10.1034/j.1399-3054.2003.00138.x

Salvo, S., Muñoz, C., Ávila, J., Bustos, J., Ramirez-Valdivia, M., Silva, C., et al. (2012). An estimate of potential blueberry yield using regression models that relate the number of fruits to the number of flower buds and to climatic variables. *Sci. Hortic.* 133, 56–63. doi: 10.1016/j.scienta.2011.10.020

Sangster, T. A., Bahrami, A., Wilczek, A., Watanabe, E., Schellenberg, K., McLellan, C., et al. (2007). Phenotypic diversity and altered environmental plasticity in *Arabidopsis thaliana* with reduced Hsp90 levels. *PLoS ONE* 2:e648.doi: 10.1371/journal.pone.0000648

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682. doi: 10.1038/nmeth.2019

Schulz, D. F., Schott, R. T., Voorrips, R. E., Smulders, M. J., Linde, M., and Debener, T. (2016). Genome-wide association analysis of the anthocyanin and carotenoid contents of rose petals. *Front. Plant Sci.* 7:1798. doi: 10.3389/fpls.2016.01798

Sharpe, R. H., and Darrow, G. M. (1959). Breeding blueberries for the Florida climate. *Proc. Fla. State Hort. Soc.* 72, 308–311.

Sillanpää, M. J. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* 106, 511–519. doi: 10.1038/hdy.2010.91

Singh, R., Rastogi, S., and Dwivedi, U. N. (2010). Phenylpropanoid metabolism in ripening fruits. *Compr. Rev. Food Sci. Food Saf.* 9, 398–416. doi: 10.1111/j.1541-4337.2010.00116.x

Slater, A. T., Cogan, N. O., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2016.02.0021

Slater, A. T., Wilson, G. M., Cogan, N. O., Forster, J. W., and Hayes, B. J. (2013). Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. *Theor. Appl. Genet*. 127, 809–820. doi: 10.1007/s00122-013-2258-7

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Spoelhof, J. P., Soltis, P. S., and Soltis, D. E. (2017). Pure polyploidy: closing the gaps in autopolyploid research. *J. Syst. Evol.* 55, 340–352. doi: 10.1111/jse.12253

Stone, S. L. (2014). The role of ubiquitin and the 26S proteasome in plant abiotic stress signaling. *Front. Plant Sci.* 5:135. doi: 10.3389/fpls.2014.00135

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genome-wide experiments. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Tailor, S., Bykova, N. V., Igamberdiev, A. U., and Debnath, S., C. (2017). Structural pattern, and genetic diversity in blueberry (*Vaccinium*) clones and cultivars using EST-PCR and microsatellite markers. *Genet. Resour. Crop Evol.* 64, 1–12. doi: 10.1007/s10722-017-0497-1

Tamayo-Ordóñez, M. C., Espinosa-Barrera, L. A., Tamayo-Ordóñez, Y. J., Ayil-Gutiérrez, B., and Sánchez-Teyer, L. F. (2016). Advances and perspectives in the generation of polyploid plant species. *Euphytica* 209, 1–22. doi: 10.1007/s10681-016-1646-x

Tanksley, S. D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *Plant Cell* 16(Suppl.), S181–S189. doi: 10.1105/tpc.018119

Teyssier, E., Bernacchia, G., Maury, S., How Kit, A., Stammitti-Bert, L., Rolin, D., et al. (2008). Tissue dependent variations of DNA methylation and endoreduplication levels during tomato fruit development and ripening. *Planta* 228, 391–399. doi: 10.1007/s00425-008-0743-z

Thompson, E. P., Llewellyn Smith, S. G., and Glover, B. J. (2012). An Arabidopsis rhomboid protease has roles in the chloroplast and in flower development. *J. Exp. Bot*. 63, 3559–3570. doi: 10.1093/jxb/ers012

Uitdewilligen, J. G., Wolters, A.-M. A., Bjorn, B., Borm, T. J. A., Visser, R. G. F., and van Eck, H. J. (2015). Correction: A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* 8:e0141940. doi: 10.1371/journal.pone.0141940

USDA (2016). *United States Department of Agriculture: Fruit and Tree Nut Data*. Avaiable online at: https://data.ers.usda.gov.

Van de Poel, B., Bulens, I., Oppermann, Y., Hertog, M. L., Nicolai, B. M., Sauter, M., et al. (2013). S-adenosyl-l-methionine usage during climacteric ripening of tomato in relation to ethylene and polyamine biosynthesis and transmethylation capacity. *Physiol. Plant.* 148, 176–188. doi: 10.1111/j.1399-3054.2012.01703.x

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wang, W., Cai, J., Wang, P., Tian, S., and Qin, G. (2017). Post-transcriptional regulation of fruit ripening and disease resistance in tomato by the vacuolar protease SlVPE3. *Genome Biol.* 18:47. doi: 10.1186/s13059-017-1178-2

Wang, Y., Wang, W., Cai, J., Zhang, Y., Qin, G., and Tian, S. (2014). Tomato nuclear proteome reveals the involvement of specific E2 ubiquitin-conjugating enzymes in fruit ripening. *Genome Biol.* 15:548. doi: 10.1186/s13059-014-0548-2

Wang, Y., Zhang, W. Z., Song, L. F., Zou, J. J., Su, Z., and Wu, W. H. (2008). Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. *Plant Physiol.* 148, 1201–1211. doi: 10.1104/pp.108.126375

Würschum, T., Liu, W., Gowda, M., Maurer, H. P., Fischer, S., Schechert, A., et al. (2012). Comparison of biometrical models for joint linkage association mapping. *Heredity* 108, 332–340. doi: 10.1038/hdy.2011.78

Zhang, Z., Jiang, S., Wang, N., Li, M., Ji, X., Sun, S., et al. (2015). Identification of differentially expressed genes associated with apple fruit ripening and softening by suppression subtractive hybridization. *PLoS ONE* 10:e0146061. doi: 10.1371/journal.pone.0146061

Conflict of Interest Statement: JO and CC were affiliated to the University of Florida when the study started and by the time the manuscript was submitted they were employed by Driscoll's Inc., and Duda Farm Fresh Foods, respectively.

The remaining 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: 03 July 2018
doi: 10.3389/fpls.2018.00944

Genome Reduction in Tetraploid Potato Reveals Genetic Load, Haplotype Variation, and Loci Associated With Agronomic Traits

*Norma C. Manrique-Carpintero[1], Joseph J. Coombs[1], Gina M. Pham[2], F. Parker E. Laimbeer[3], Guilherme T. Braz[2], Jiming Jiang[2,4], Richard E. Veilleux[3], C. Robin Buell[2,5] and David S. Douches[1]**

[1] Potato Breeding and Genetics Program, Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, United States, [2] Department of Plant Biology, Michigan State University, East Lansing, MI, United States, [3] Department of Horticulture, Virginia Tech, Blacksburg, VA, United States, [4] Department of Horticulture, Michigan State University, East Lansing, MI, United States, [5] Plant Resilience Institute, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:
*Richard John Abbott,
University of St Andrews,
United Kingdom*

Reviewed by:
*Laura M. Shannon,
University of Minnesota Twin Cities,
United States
Ek Han Tan,
University of Maine, United States*

***Correspondence:**
*David S. Douches
douchesd@msu.edu*

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

Received: 15 March 2018
Accepted: 12 June 2018
Published: 03 July 2018

Citation:
Manrique-Carpintero NC, Coombs JJ, Pham GM, Laimbeer FPE, Braz GT, Jiang J, Veilleux RE, Buell CR and Douches DS (2018) Genome Reduction in Tetraploid Potato Reveals Genetic Load, Haplotype Variation, and Loci Associated With Agronomic Traits. Front. Plant Sci. 9:944. doi: 10.3389/fpls.2018.00944

The cultivated potato (*Solanum tuberosum*) has a complex genetic structure due to its autotetraploidy and vegetative propagation which leads to accumulation of mutations and a highly heterozygous genome. A high degree of heterozygosity has been considered to be the main driver of fitness and agronomic trait performance in potato improvement efforts, which is negatively impacted by genetic load. To understand the genetic landscape of cultivated potato, we constructed a gynogenic dihaploid ($2n = 2x = 24$) population from cv. Superior, prior to development of a high-density genetic map containing 12,753 single nucleotide polymorphisms (SNPs). Common quantitative trait loci (QTL) were identified for tuber traits, vigor and height on chromosomes 2, 4, 7, and 10, while specific QTL for number of inflorescences per plant, and tuber shape were present on chromosomes 4, 6, 10, and 11. Simplex rather than duplex loci were mainly associated with traits. In general, the Q allele (main effect) detected in one or two homologous chromosomes was associated with lower mean trait values suggesting the importance of dosage allelic effects, and the presence of up to two undesired alleles in the QTL region. Loss of heterozygosity has been associated with a lower rate of fitness, yet no correlation between the percent heterozygosity and increased fitness or agronomic performance was observed. Based upon linkage phase, we reconstructed the four homologous chromosome haplotypes of cv. Superior. revealing heterogeneity throughout the genome yet nearly duplicate haplotypes occurring among the homologs of particular chromosomes. These results suggest that the potentially deleterious mutations associated with genetic load in tetraploid potato could be mitigated by multiple loci which is consistent with the theory that epistasis complicates the identification of associations between markers and phenotypic performance.

Keywords: haplotype, dihaploid, genetic load, complex traits, linkage map

INTRODUCTION

Cultivated potato is an autotetraploid, highly heterozygous, and vegetatively propagated species. Tetrasomic inheritance comprises multiple genotypic configurations with up to four alleles and various combinations of alleles and dosage per locus. The more diverse alleles are at a locus, the greater the heterozygosity and number of allelic and epistatic interactions (Carputo and Frusciante, 2011). At any given locus of a tetraploid clone, there are up to three types of intra-locus interactions that could result in non-additive effects: first order (between two alleles), second order (among three alleles), and third order (among four alleles) while allelic dosage could mediate additive effects of intra-locus interactions. There may be more complexity for the optimal allelic combinations, locus interactions, and genetic effects when modeling quantitative traits. Elevated heterozygosity and genetic load have also been considered the main drivers of high and low vigor, respectively, associated with agronomic trait performance of cultivated potato. Inbreeding depression after self-pollination, and the superiority of tetraploid potato due to heterozygosity and polyploidy established a breeding bias toward increased heterotic diversity (De Jong and Rowe, 1971; Mendoza and Haynes, 1974). Besides the effects of epistasis, gain or loss of allelic diversity could be responsible for heterosis and inbreeding depression, respectively; recessive undesired alleles in the homozygous state would be expected to decrease fitness whereas allelic diversity at heterozygous loci facilitate both dominant and overdominant effects (Miranda Filho, 1999; Ceballos et al., 2015). Consistent with this hypothesis, a large-scale genome resequencing survey of genetic load in asexually propagated cassava revealed that the amount of deleterious mutations is greater in cultivated cassava compared to wild progenitors, as has been found in maize, sunflower and rice, and that cultivated cassava has a markedly greater number of mutations in the heterozygous rather than the homozygous state which could mask the lethal effects of recessive deleterious mutations (Ramu et al., 2017; Wang et al., 2017). The asexual propagation and polyploidy of cultivated potato give the potential of retaining greater mutational load, and also the generation of genome plasticity that enhances adaption to environmental changes. Copy number variation (CNV) in cultivated tetraploid potato is widely distributed throughout the genome and has been associated with lowly expressed genes and genes that respond to biotic and abiotic stress (Pham et al., 2017).

Dihaploids ($2n = 2x = 24$) from the cultivated tetraploid potato *Solanum tuberosum* ($2n = 4x = 48$) have been a valuable tool for genetic and cytogenetic studies as well as for breeding. Peloquin et al. (1991) reviewed the use of dihaploids to support evidence of tetrasomic inheritance, determine the basic chromosome number within the *Solanum* genus, discover meiotic mutations, understand ploidy and evolution, and assess sexual compatibility and hybridization barriers in potato. Dihaploid progeny of potato can be produced by anther culture or by chromosome elimination, sometimes referred to as "prickle pollination." Specific haploid-inducer lines induce chromosome elimination; following fertilization from a cross of a tetraploid maternal clone with a haploid inducer, the paternal chromosomes are selectively eliminated from the developing hybrid embryo. By introduction of a homozygous, dominant embryo spot marker into haploid inducers, gynogenic dihaploid seed can be selected by the absence of the purple embryo spot visible on the hypocotyl of embryos or seedlings (Hermsen and Verdenius, 1973). As gametic genotypic representations of autotetraploid potato, dihaploid populations can facilitate determination of the complex genetic structure of cultivated potato. The reduced genome complexity of dihaploids enables simpler segregation ratios than tetraploids and a better understanding of the genetic factors controlling traits of interest. Several dihaploid populations have been used to decipher monogenic or polygenic inheritance and gene action effects associated with morphologic, agronomic and disease resistance traits (Cipar and Lawrence, 1972; Matsubayashi, 1979; De Maine, 1984; Pineda et al., 1993; Song et al., 2005; Velasquez et al., 2007). Unilateral ($4x \times 2x$) and bilateral ($2x \times 2x$) crosses using dihaploids have also served as a bridge to generate simpler and more efficient breeding schemes, overcome hybridization barriers, and achieve introgression of adaptive traits in the cultivated potato (Chase, 1963; Peloquin et al., 1991; Rokka, 2009).

The potato cultivar "Superior" was released in 1962 by the University of Wisconsin as a round white variety with scab resistance, and medium maturity (Rieman, 1962). Currently, it is grown in the USA and Canada as a fresh market variety. Dihaploid populations of a potato variety exhibit uniparental segregation following genome reduction. Using cv. Superior. as a model in our study, we generated a dihaploid population extracted from cv. Superior. to observe the effects of unmasking genetic load on different agronomic traits and elucidate main genomic regions associated with trait performance to understand the genetic complexity of tetraploid potato.

MATERIALS AND METHODS

Plant Material

A gynogenic dihaploid ($2n = 2x = 24$) population of 95 individuals was created from *S. tuberosum* Group Tuberosum tetraploid cv. Superior. The *S. tuberosum* Group Phureja haploid inducer IVP101, homozygous dominant for an embryo seed spot marker, was used as the pollinator. Seeds lacking a purple spot were grown and leaf tissue from *in vitro* plantlets subjected to flow cytometry to identify dihaploids (Owen et al., 1988). Peaks were compared to known monoploid and diploid controls.

Genotyping

DNA was isolated from leaf tissue of the "Superior" parent and 95 dihaploid gynogenic progeny and Illumina compatible paired end libraries were constructed as described previously (Hardigan et al., 2016). Libraries were skim-sequenced on the Illumina HiSeq 2000 platform at low coverage, with a theoretical approximation of 8x coverage of the genome, to identify single nucleotide polymorphic (SNP) segregating markers. Adapters and low quality bases were removed from the raw reads using Cutadapt v. 1.8.1 (Martin, 2011) and cleaned reads were aligned using BWA-MEM v. 0.7.11r1034 (Li, 2013) to the *S. tuberosum* Group Phureja DM 1-3 516 R44 reference genome v4.04 (Hardigan et al., 2016). Genotypes were called using the GATK Unified Genotyper (McKenna et al., 2010). Markers with unexpected segregation, distorted segregation (Chi-square threshold P-value <0.01), and singleton markers (SNPs without any duplicates) or markers with just one duplicate were removed. The remaining high quality markers were used for map construction after excluding duplicate co-segregating markers. Raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA335821.

Linkage Map and Quantitative Trait Locus Analysis

The TetraploidSNPMap software for biallelic SNP markers (Hackett et al., 2017), informative for allele dosage in an autotetraploid species, was used to generate the genetic map and quantitative trait locus (QTL) analysis. As a unique parent population, only simplex (AAAB, ABBB) and duplex (AABB) marker configurations in "Superior" were segregating in the dihaploid population. The expected segregation for simplex markers in the diploid progeny corresponded to a 1:1 homozygous: heterozygous genotypic ratio (AA:AB, BB:AB), and for duplex markers to a 1:4:1 genotypic ratio (AA:AB:BB). However, four homologs per parental chromosome are segregating in this population. Thus, the segregation obtained in the dihaploid progeny fits the autotetraploid segregation for a cross with a null male parent for simplex (AAAB × AAAA, ABBB × BBBB) and duplex (AABB × AAAA, AABB × BBBB) markers. The marker configurations of the different genotypes were recoded according to TetraploidSNPMap code (AAAA = 0, AAAB = 1, AABB = 2, ABBB = 3, and BBBB = 4). For simplex segregation (AA:AB = AAAA:AAAB, BB:AB = BBBB:ABBB), genotypes were recoded as 0 and 1; while for duplex segregation (AA:AB:BB = AAAA:AABB:BBBB), genotypes were recoded as 0, 1, and 2.

The linkage map was constructed according to Hackett et al. (2013). The different mapping steps were implemented in TetraploidSNPMap: analysis of single marker segregation; cluster into linkage groups; estimation of recombination frequency and logarithm_{10} of the odds ratio for linkage (LOD score); and ordering and inference of SNP linkage phase (Hackett et al., 2017). A preliminary test of cluster of simplex SNPs was done using JoinMap 4.1 (Van Ooijen, 2006). Markers were coded for cross-pollinated population type (<lmxll>). This step allowed identification and exclusion of problematic markers that did not cluster as part of linkage groups. In the mapping process in TetraploidSNPMap, problematic and near duplicate markers were also detected and excluded; these mainly corresponded to outliers in the clustering and metric multidimensional scaling (MDS) ordering steps. A high concordance between genetic and physical maps has been reported for potato mapping populations (Felcher et al., 2012; Sharma et al., 2013). As a final quality control of the generated linkage maps, marker genetic positions (cM) were plotted against their physical positions (Mb) on each chromosome to generate MaryMaps (Chakravarti, 1991).

Square root transformation of phenotypic data was performed to improve the QTL detection. A QTL interval mapping analysis with a step size of 1 cM was done to identify QTL. A logarithm of the odds (LOD) threshold calculated based on a test of 500 permutations was used to detect significant marker associations. Next, the trait was modeled as an additive function of the QTL allele effect on each of eight homologous chromosomes (four for each parent). In addition to a full additive model, any of four different QTL simple models could be fit in this population with a single parent segregation. Simplex QTL model (Qqqq × qqqq), where the Q allele drives the main effect, duplex QTL (QQqq × qqqq) with additive effects of Q allele (the QTL genotypes qqqq, Qqqq, QQqq have means of m, m+Q, m+2Q), duplex QTL with non-additive effects of Q allele (the QTL genotypes qqqq, Qqqq, QQqq have different means m1, m2, m3), and duplex QTL with dominant effects of Q allele (two QTL genotypes qqqq, Q_qq mean categories). A QTL fit a simple model when the value of Schwarz Information Criterion (SIC) (Schwarz, 1978) was smaller than or close to the value of the full model, at least with a difference of 2 units from other simple models.

Field Evaluations

The dihaploid population was grown from greenhouse- or field-produced tubers at the Montcalm Research Center, Lakeview, MI (MRC) and the Botany and Plant Pathology Farm, East Lansing, MI (BPP) of Michigan State University over 2 years. In 2014, greenhouse-grown tubers were harvested in March and planted at MRC. In 2015, tubers produced in the field in 2014 in addition to greenhouse-grown tubers, harvested between February and March, were planted at MRC and BPP, respectively. In 2015, greenhouse tubers were subjected to a Rindite treatment to break dormancy prior to planting (Varga and Ferenczy, 1956). Thus, a total of three location/year datasets is reported in this study. All trials had a randomized complete block design using plots of eight plants as experimental unit and three replications per clone. Parents and progeny were evaluated for eight traits: total tuber yield (TTY) measured as g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant vigor (Vigor) scored as overall plant canopy development ∼3 months after planting using a 1–5 scale (1: low vigor, 5: high vigor), plant height (Height) in cm assessed when plants started flowering, number of inflorescences per plant counted after a line initiated flowering in the plot (Infl/plant), specific gravity (SPGR) calculated using the formula [air weight/(air weight–water weight)] for a minimal sample size of 1 kg/plot, and tuber shape (Shape) scored using a 1–5 scale (1 = compressed, 2 = round, 3 = oval, 4 = oblong and 5 = long).

Heritability and Correlation Analysis

The restricted maximum likelihood method (REML) was used to calculate broad-sense heritability (H^2) with clones as random effects and site-year environments as random fixed effects. The heritability was estimated on a genotype mean basis as the ratio of:

$$H^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \frac{\sigma_{g*s-y}^2}{m} + \frac{\sigma_e^2}{rm})}$$

where (σ_g^2), $(\frac{\sigma_{g*s-y}^2}{m})$, and $(\frac{(\sigma_e^2)}{rm})$ are the genetic, genotype × site-year environment interaction and residual variance components, m is the number of site-year environments and r is the number of replications.

Pearson correlation was used to estimate correlations between traits among site-year environments using the REML method when samples were missing. Means, variances, correlation and distribution analyses were calculated using JMP® 10 SAS Institute Inc., Cary, NC, USA.

Fluorescence *in Situ* Hybridization

Chromosome preparation and fluorescence *in situ* hybridization (FISH) were performed using published protocols (Braz et al., 2018). Individual potato chromosomes of "Superior" were identified using two "barcode probes," which contain 27,306 and 27,366 oligonucleotides (45 nt), respectively, derived from 26 different regions on the 12 potato chromosomes. These two probes produce 26 distinct FISH signals. Each of the 12 potato chromosomes is labeled with distinct signal pattern (Braz et al., 2018). FISH images were captured using a QImaging Retiga EXi Fast 1394 CCD camera and were processed with Meta Imaging Series 7.5 software. The final contrast of the images was processed using Adobe Photoshop CS3 software.

Rescue of Dwarf Mutant With Gibberellic Acid Treatment

Plantlets of the dihaploid VT_SUP_46 from the "Superior" dihaploid population maintained *in vitro*, were obtained after subculture on regular MS medium (Murashige and Skoog, 1962) (MS basal medium with vitamins + 3% sucrose + 0.6% plant agar micropropagation grade, pH 5.8, reagents from Phytotechnology Laboratories, Shawnee Mission, KS, USA). Assay tubes with plantlets were placed in a growth room at 22°C and 16-h photoperiod. Plantlets were grown in regular MS medium and medium supplemented with 0.3 mg/l of zeatin riboside (ZR, *trans* isomer, Sigma, St Louis MO USA) and two concentrations of gibberellic acid (GA_3, Research Products International, Mt Prospect, IL, USA), 0.02 or 0.2 mg/l to rescue from the unique dwarf phenotype observed in this clone.

RESULTS

Phenotypic Performance

A total of 95 dihaploid clones was generated from crosses of cv. Superior. with IVP101; however, due to the low vigor of many of the dihaploids there was limited production of planting material in the greenhouse, and/or delayed emergence in the field such that between 50 and 75 individual clones were evaluated for the various traits under field conditions. For SPGR, 39 clones could be assessed in the MRC-2014 trial (**Table 1**). A wide range of variation of traits was observed in the population (**Table 1**). The quantile diagnostic plots showed a trend toward normal distribution for the population means of TS, Height, SPGR, and Shape, while bimodal normal distributions were observed for the means of TTY, ATW, Vigor and Infl/plant (Supplementary Figure 1). Transgressive segregation for TS, Vigor, Height, Infl/plant, and SPGR was detected in the progeny. For TTY and ATW, a few progeny performed similar to the parental line skewing the distribution toward greater values. Similarly, a few individuals with many Infl/plant skewed the distribution of this trait, especially in MRC-2014 and BPP-2015 site-year environments. Even though the normal distribution for SPGR was not affected by the skewness, there was tendency toward low values as shown by the negative skewness.

Overall, the number of days after planting required for 75% of plants per clone in a plot to emerge varied from 24 to 100 days. MRC-2014 had a significantly greater average number of days (63) to emerge compared with 52 and 42 days for BPP-2015 and MRC-2015, respectively (*P*-value <0.0001). The high correlation between MRC-2014 and BPP-2015 emergence data (0.53, *P*-value <0.0001), and low correlation of these locations with MRC-2015 (0.36 and 0.30, respectively; *P*-value < 0.01 and 0.02, respectively) showed that the greenhouse-produced tubers used as planting material at both locations had similar longer emergence period compared to the field-grown tubers from the previous season that was used as seed for MRC-2015; this seed source appeared to be the main driver of more efficient emergence for most of the dihaploid clones (35.9 days for 75% of the population). The planting material did not have a critical effect on the reproducibility of data as the broad sense heritability was greater than 0.7 for all traits (**Table 2**). Comparison of correlations among data from different locations for the same trait, revealed that nearly all correlations were greater than 0.5 with *P*-values <0.0001. A low correlation was observed only for SPGR in BPP-2015 with MRC-2014 and MRC-2015 (0.24 and 0.4; *P*-values < 0.01 and 0.004, respectively), whereas the correlation between MRC-2014 and MRC-2015 for SPGR was 0.85 (*P*-value <0.0001). The low tuber yield for many individuals limited the total number of progeny evaluated for SPGR, therefore this trait was excluded from the QTL analysis.

High positive correlations among TTY, TS, ATW, Height, and Vigor were observed for all 3 site-year environments (**Tables 3–5**). Infl/plant showed high and moderate positive correlations for all 3 site-year locations. Tuber shape had low to no correlation with the other traits. The longer emergence period was highly correlated with low Height and Vigor for all three environments, while moderate to low negative correlations were observed between emergence and the three tuber traits, TTY, ATW, and TS.

A site-year environmental effect was detected for all traits except ATW. MRC-2015 reported significantly greater mean values and MRC-2014 the lowest for TS, Height, and Vigor, while BPP-2015 had the greatest values for Infl/plant and Shape, and BPP-2015 and MRC-2015 for TTY [*P*-values <0.001 for all except ATW (0.072) and Shape (0.008)]. For 53 dihaploid clones with full data, we detected significant (*P*-value <0.0001) genotype-environment interactions in all locations for TTY, TS, ATW, Vigor, Height, and Shape.

Linkage Map

A high-density genetic map was built for the 95-progeny of "Superior" dihaploid population. After filtering to identify high-quality segregating markers (Supplementary Table 1), we identified 12,753 polymorphic SNPs that were successfully

TABLE 1 | Frequency distribution statistics of cv. Superior. and its dihaploid population (Pop) for 3 site-year environments [Montcalm Research Center (MRC) in 2014 and 2015, and Botany and Plant Pathology Farm (BPP) in 2015].

Trait	Site-Year	Superior cv. Mean	Pop N	Pop Mean	Pop Std Dev	Pop Min	Pop Max	Pop Skewness	Pop % CV	Pop Kurtosis	% Inbreeding depression
TTY	MRC-2014	605	59	135	124.09	7.13	579.88	1.418	92.25	1.99	77.8
	BPP-2015	1013	75	255	251.09	1.60	1119.75	1.634	98.65	2.64	74.9
	MRC-2015	921	58	260	190.05	0.00	781.54	0.897	73.17	0.35	71.8
ATW	MRC-2014	138	59	40	19.57	11.50	109.06	1.175	49.52	1.80	71.4
	BPP-2015	165	75	42	24.94	4.00	111.84	0.671	59.27	0.02	74.5
	MRC-2015	107	58	40	19.91	0.00	100.94	0.963	49.68	0.95	62.4
TS	MRC-2014	4.36	62	2.84	1.90	0.00	7.22	0.268	66.75	−0.80	34.8
	BPP-2015	6.14	75	5.02	3.24	0.36	13.40	0.726	64.57	0.12	18.2
	MRC-2015	8.67	58	6.15	3.35	0.00	13.59	0.267	54.46	−0.54	29.1
Height	MRC-2014	49.33	61	26.41	13.59	6.00	53.67	0.176	51.45	−1.07	46.5
	BPP-2015	56.50	74	35.18	14.57	9.00	71.33	0.229	41.42	−0.44	37.7
	MRC-2015	51.33	58	38.23	14.22	14.00	71.00	0.117	37.19	−0.80	25.5
Vigor	MRC-2014	4.33	61	2.62	1.19	1.00	5.00	0.053	45.45	−1.03	39.5
	BPP-2015	4.50	75	3.13	1.21	1.00	5.00	0.005	38.71	−1.12	30.5
	MRC-2015	5.00	58	3.49	1.35	1.00	5.00	−0.526	38.55	−1.10	30.2
Inf/plant	MRC-2014	1.67	50	1.40	1.28	0.00	6.00	2.273	91.42	6.05	16.2
	BPP-2015	2.00	69	2.86	2.43	0.30	10.67	1.629	84.96	2.03	−43.1
	MRC-2015	0.80	58	1.39	1.50	0.00	5.30	1.090	107.96	0.00	−73.2
SPGR	MRC-2014	1.07	39	1.07	0.01	1.05	1.10	−0.096	1.03	−0.16	−0.4
	BPP-2015	1.07	68	1.06	0.01	1.04	1.08	−0.442	0.96	0.16	0.9
	MRC-2015	1.08	54	1.07	0.01	1.04	1.09	−0.762	0.90	0.42	0.7
Shape	MRC-2014	2.83	60	2.86	0.87	1.50	5.00	0.769	30.29	−0.02	−1.0
	BPP-2015	2.50	75	3.08	0.56	2.00	4.67	0.401	18.10	−0.12	−23.3
	MRC-2015	2.17	57	2.83	0.73	1.67	4.83	0.910	25.91	0.43	−30.6

Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 = compressed, 2 = round, 3 = oval, 4 = oblong and 5 = long.

TABLE 2 | Heritability for eight agronomic traits evaluated in the "Superior" dihaploid population.

Trait	Heritability (%)
TTY	92.3
ATW	90.8
TS	86.9
Height	83.9
Vigor	89.0
Infl/plant	77.9
SPGR	75.8
Shape	88.6

Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 = compressed, 2 = round, 3 = oval, 4 = oblong and 5 = long.

mapped (**Table 6** and Supplementary Table 2). The SNPs were located mainly at intergenic regions (10,159, 79.7%), compared to genic regions (2594, 20.3%), with 746 in exons and 1,970 in introns, and 120 overlapping both positions due to alternative splicing. The genetic map has a length of 1299.1 cM with 819 to 1374 SNPs per chromosome. The average inter-locus distance was 0.7 cM with a genome coverage of 99.3% relative to the 12 chromosomes in the current potato genome assembly.

QTL Identified

In general, common QTL were identified on chromosomes 2, 4, 7, and 10 for TTY, TS, ATW, Height, and Vigor, while specific QTL were identified for Infl/plant and Shape on chromosomes 4, 6, 10, and 11 (**Figure 1**). In some cases, the QTL were not identified in all site-year environments for each trait, as the peak was not always significant or not detected. **Table 7** summarizes the QTL chromosome locations, phenotypic variation, QTL genetic model and homologous chromosomes associated with the Q allele effect. For most of the QTL, the closest SNPs to the QTL peak, co-segregating in phase with the Q alleles, were also reported.

Fluorescence *in Situ* Hybridization

Even though a strict quality filtering process was used in the selection of markers for linkage mapping, the construction of genetic maps for chromosomes 4 and 11 was especially problematic. We conducted oligo-based fluorescence *in situ* hybridization (Oligo-FISH) (Braz et al., 2018) to examine if the four copies of chromosomes 4 and 11 in "Superior"

TABLE 3 | Correlation analysis for field season at Montcalm Research Center in 2014 for nine traits.

	TTY	ATW	TS	Height	Vigor	SPGR	Infl/pant	Shape	75% Emerg
TTY	1								
ATW	0.87	1							
TS	0.80	0.51	1						
Height	0.82	0.74	0.73	1					
Vigor	0.83	0.74	0.78	0.94	1				
SPGR	0.06	−0.06	0.16	−0.15	−0.11	1			
Infl/plant	0.59	0.46	0.48	0.63	0.68	0.10	1		
Shape	0.20	0.29	0.12	0.33	0.29	−0.08	0.40	1	
75% Emerg	− 0.40	− 0.29	− 0.42	− 0.63	− 0.63	0.38	− 0.42	−0.15	1

Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 = compressed, 2 = round, 3 = oval, 4 = oblong and 5 = long, 75% of emergence number of days after planting (75% Emerg). Significant positive correlation dark green P-value < 0.0001, light green P-value < 0.05, significant negative correlation dark red P-value < 0.0001, intermediate red P-value < 0.001, and light red P-value < 0.001.

TABLE 4 | Correlation analysis for field season at Botany and Plant Pathology Farm in 2015 for nine traits.

	TTY	ATW	TS	Height	Vigor	SPGR	Infl/pant	Shape	75% Emerg
TTY	1								
ATW	0.84	1							
TS	0.80	0.48	1						
Height	0.74	0.70	0.61	1					
Vigor	0.74	0.71	0.66	0.93	1				
SPGR	0.44	0.47	0.37	0.48	0.52	1			
Infl/plant	0.60	0.58	0.47	0.74	0.72	0.23	1		
Shape	0.02	0.20	−0.16	0.07	0.13	0.005	0.23	1	
75% Emerg	− 0.39	− 0.40	− 0.30	− 0.57	− 0.62	− 0.34	− 0.27	−0.08	1

Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 = compressed, 2 = round, 3 = oval, 4 =oblong and 5 = long, 75% of emergence number of days after planting (75% Emerg). Significant positive correlation dark green P-value < 0.0001, light green P-value < 0.05, significant negative correlation dark red P-value < 0.0001, intermediate red P-value < 0.001, and light red P-value < 0.001.

show visible structural variation. All 48 chromosomes could be individually identified based on the Oligo-FISH signal patterns (**Figure 2**). We did not observe any unambiguous chromosome structural changes associated with chromosomes 4 and 11. However, three copies of chromosome 4 contain a visible heterochromatic knob in the short arm, whereas the remaining copy of chromosome 4 does not contain the knob (**Figure 2**).

Double Reduction Leads to a Dwarf Mutant

A dark green and rosette dwarf phenotype that can be rescued by GA_3 application has been reported in hybrid progeny of cv. Superior. as well as some other potato diploid and tetraploid clones (Bamberg and Hanneman, 1991; Valkonen et al., 1999). A single dihaploid, VT_SUP_46, within our "Superior" dihaploid population has a strong dwarf phenotype (**Figure 3**). Treatment of *in vitro* plantlets of VT_SUP_46 on propagation medium supplemented with GA_3 (0.02 and 0.2 mg/l) resulted in rescue from the dwarf phenotype (**Figure 3**).

DISCUSSION

Genetic Load Unmasked in cv. Superior. Dihaploid Population

As reported previously (Peloquin and Hougas, 1960; De Maine, 1984; Kotch et al., 1992; Hutten et al., 1995), segregation of a tetraploid parent configuration in a gametic dihaploid population leads to breakdown of allelic combinations and interactions, and to unmasking of the genetic load due to homozygosity of recessive alleles and/or the effects of dysfunctional alleles. A dihaploid population has an expected reduction of heterozygosity equivalent to three generations of self-pollination of an autotetraploid, which increases the probability of a homozygous state of recessive and deleterious alleles (Peloquin and Hougas, 1960). The effect of homozygous recessive and sub-lethal alleles in a duplex configuration in a locus in the parental line will lead to 17% weakness or loss in the progeny, to 50% when in a triple dose in a triplex parent genotype, and would not be detected in a simplex configuration (Hutten et al., 1995). In complex traits, several genes and their contribution to the genetic structure of the

TABLE 5 | Correlation analysis for field season at Montcalm Research Center in 2015 for nine traits.

	TTY	ATW	TS	Height	Vigor	SPGR	Infl/pant	Shape	75% Emerg
TTY	1								
ATW	0.74	1							
TS	0.77	0.26	1						
Height	0.65	0.56	0.51	1					
Vigor	0.70	0.54	0.65	0.91	1				
SPGR	0.12	0.10	−0.03	0.09	0.04	1			
Infl/plant	0.46	0.40	0.34	0.80	0.68	0.15	1		
Shape	0.11	0.38	−0.04	0.20	0.22	−0.07	0.27	1	
75% Emerg	−0.35	−0.26	−0.39	−0.44	−0.53	0.006	−0.18	−0.08	1

Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 = compressed, 2 = round, 3 = oval, 4 = oblong and 5 = long, 75% of emergence number of days after planting (75% Emerg). Significant positive correlation dark green P-value < 0.0001, light green P-value < 0.05, significant negative correlation dark red P-value < 0.0001, intermediate red P-value < 0.001, and light red P-value < 0.001.

TABLE 6 | "Superior" linkage map length in centimorgans (cM), physical length in megabase pairs (Mb), and features of mapped single nucleotide polymorphisms (SNPs).

							Interval distance cM		
Chr	**Total SNP**	**N Seg**	**Bins**	**cM**	**Mb**	**PGSC v4.03 Mb**	**Min**	**Max**	**Mean**
chr01	1216	182	170	108.8	88.3	88.7	0.01	5.98	0.64
chr02	853	174	166	106.8	47.1	48.6	0.01	6.43	0.65
chr03	1235	169	165	106.4	62.01	62.3	0.01	4.89	0.65
chr04	1148	153	129	143.8	72.01	72.2	0.01	9.02	1.12
chr05	819	169	154	96.9	51.8	52.1	0.01	4.42	0.63
chr06	1246	164	158	92.5	58.9	59.5	0.01	3.95	0.59
chr07	1374	178	170	94.8	56.6	56.8	0.01	3.69	0.56
chr08	1201	139	139	90.3	56.7	56.9	0.01	5.37	0.65
chr09	847	149	143	104	61.3	61.5	0.01	4.73	0.73
chr10	978	119	109	91.4	59.5	59.8	0.01	5.43	0.85
chr11	897	134	116	159.8	45.2	45.5	0.01	19.88	1.39
chr12	939	210	187	103.7	60.8	61.2	0.01	6.12	0.56
Total	12753	1940	1806	1299.1	720.1	725.1	0.01	19.88	0.72

Chromosome (Chr), potato genome sequence consortium assembly version 4.03 (PGSC v 4.03), unique segregating markers (Seg).

trait will influence the magnitude of the effect of recessive or sub-lethal alleles, producing a wide range of variation in phenotype. Hutten et al. (1995) evaluated 31 different dihaploid populations reporting some levels of dwarfism, wide variation between populations in the rate of tuberization ability, and low frequencies of flowering and pollen stainability. In fact, low fitness phenotypes prevented 35.8% (MRC-2014), 22.1% (BPP-2015), and 40% (MRC-2015) of the "Superior" dihaploid population to be evaluated under different site-years. Almost 20% of "Superior" dihaploids were never evaluated in the field due to extremely low vigor (**Figure 4**).

Using the theory proposed by Fasoulas (1988) where greater genetic load affecting a trait would increase the coefficient of variation (%CV), causing negative kurtosis and positively skewing the trait frequency distribution of a dihaploid population, Kotch et al. (1992) studied the frequency distribution statistics of several dihaploid potato populations. Skewness, kurtosis, and the inbreeding depression coefficient (relative percentage of dihaploid population mean compared to tetraploid parent mean) were used to indicate the type of gene action affecting different traits. In general, a negligible or low inbreeding depression coefficient, close to zero or minor skewness, negative or zero kurtosis, and low %CV indicated that the 4*x* parent had genes with primarily additive effects and low genetic load associated with the trait. In contrast, significant positive skewness, positive kurtosis, high %CV, and a high inbreeding depression coefficient are suggestive that the 4*x* parent primarily has genes with non-additive effects associated with the trait. In this "Superior" dihaploid population (**Table 1**), SPGR is potentially a trait mainly governed by genetic factors with additive effects and low genetic load, while for TTY and ATW genes that have non-additive effects and greater genetic load are suggested. For traits where the distribution statistics fit in the middle of these parameters, both additive and non-additive

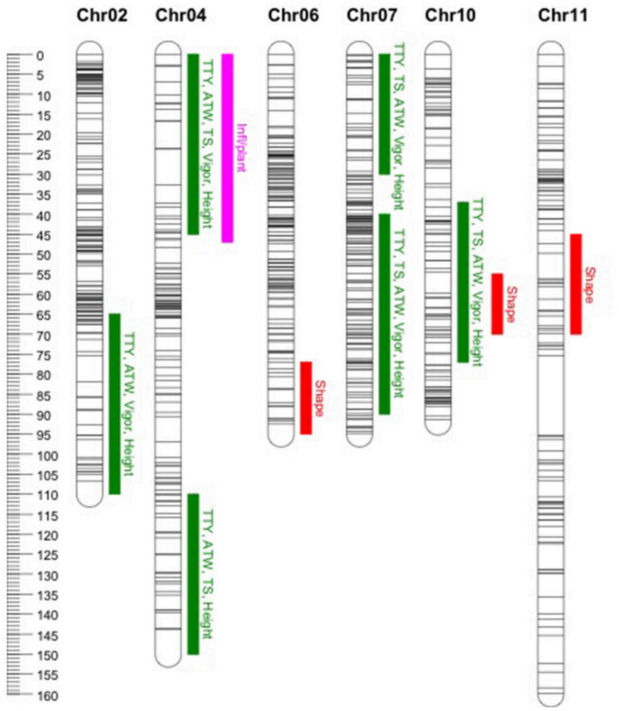


FIGURE 1 | Location of identified quantitative trait loci (QTL). Common QTL as listed in **Table 7** for total tuber yield (TTY), tuber set (TS), average tuber weight (ATW), Vigor and Height (green). Specific QTL for number of inflorescences per plant (Inf/plant) and tuber shape (Shape) in pink and red, respectively. On the left the cM genetic position scale for the mapped SNPs represented with horizontal lines along the chromosomes.

genetic effects could be mediating the phenotype, this is the case for TS, Height, Vigor, and Infl/plant.

By comparing the performance of different populations, Kotch et al. (1992) highlighted that a trait with similar %CV could have different inbreeding depression coefficients, which implies the importance of non-additive gene control rather than genetic load (fixation of deleterious genes). Evaluations of trait performance in inbred generations and diallelic crosses of outcrossing species (e.g., maize, cassava) suggest that the relevance of non-additive effects increases with the genetic complexity of a trait, and that a strong inbreeding depression effect will also be associated (Ceballos et al., 2015). Non-additive effects driving heterosis (dominance, overdominance, and epistasis) are particularly important for grain yield and fresh root/tuber yield. A similar complexity is suggested in potato yield traits, such as tuber yield with strong inbreeding depression as described in this analysis and reported in populations of self-pollinated tetraploids (Golmirzaie et al., 1998).

Genome Heterogeneity of cv. Superior

A high-density genetic linkage map was built for the 95-progeny of "Superior" dihaploid population. For several chromosomes, it was difficult to order and estimate the linkage phase of the markers, particularly for chromosomes 4 and 11. With the MDS ordering approach (Preedy and Hackett, 2016), a continuous curve plot is expected because of the low linkage between markers located at opposite chromosome ends. The MDS graph showed sub-clusters of markers producing extra sub-curves within the general curve or outlier points in some instances. Excluding problematic markers solved this problem. Primarily genotype errors or distorted segregation could affect the marker quality and mapping process. This is not the case in our population since besides the threshold (P-value < 0.01) used to eliminate markers with distorted segregation, we did not detect any pattern with meaningful distorted segregation that could limit transmission of specific genomic regions. However, inversions in some homologs or structural variation between homologous chromosomes could also be associated with problems during linkage mapping. In fact, an inversion resulted in a large gap on chromosome 11, while on chromosome 4 we observed a tendency of independent clustering and mapping of the homologous chromosomes. Both chromosomes 4 and 11 showed a greater length than normally reported in previous diploid and tetraploid linkage maps (Hackett et al., 2013; Sharma et al., 2013; Manrique-Carpintero et al., 2015; Massa et al., 2015; Da Silva et al., 2017). For chromosome 4, this may be due to the large heterochromatic knob on three of the four homologous chromosomes (**Figure 2**).

Based on the linkage phase generated in the mapping process, we reconstructed the four haplotypes for each of the 12 homologous chromosomes of the "Superior" parent for the mapped loci (**Figure 5**). Then genetic distances were calculated between different pairs of homologs per chromosome using GGT 2.0 software (Van Berloo, 2008). Different patterns of differentiation among homologs per chromosome were observed based on the simple matching coefficient (the number of shared alleles as proportion of all alleles) distance measurement (**Table 8**). For instance, for chromosomes 10 and 12, only one homolog was markedly different from the other three, while for chromosomes 1, 2, and 5, a pair of homologous chromosomes were highly similar and the other types of homologs were distant. This analysis revealed a novel observation of a high level of heterogeneity among homologous chromosomes in a tetraploid potato cultivar.

QTL Analysis of Agronomic Traits

Highly correlated traits shared QTL with similar positions and effects. For most of the QTL (chromosome 2, 4, and 7) for TTY, ATW, TS, Height, and Vigor, the Q allele was in simplex configuration and associated with lower trait mean values in heterozygous genotypes. When the Q allele was detected on two homologous chromosomes, the presence of any or both Q alleles was associated with lower mean values. This resulted in having a marker segregation in which 50 or 16.7% of the evaluated population showed lower fitness phenotype with the Q allele associated. This could be explained by the importance of dosage allelic effect in the genotype configuration of the tetraploid parent or that the tetraploid parent has mainly one and up to two weak or dysfunctional alleles in the QTL regions. Nevertheless, if the recessive detrimental alleles are in simplex configuration we do not expect homozygous allelic states unless

TABLE 7 | QTL identified in the “Superior” dihaploid mapping population, chromosome (Chr), and genetic centimorgan position (cM), logarithm of the odds (LOD) significance, variance explained (R^2).

Trait	QTL locus	Chr	cM	LOD	% R2	QTL genetic model	Homologous chromosome	Site-year location	Main Q effect
TTY	chr02_47.91_11358	chr02	105	3.7	12.1	Simplex	H4	MRC-2014	↓
	chr02_47.47_d1954	chr02	102	2.6	6.2	Duplex, no additive	F14	MRC-2015	↓
	chr04_1.5_s247	chr04	9	2.5	7.7	Simplex	H2	MRC-2014	↓
	chr04_1.5_s247	chr04	10	3.1	10.1	Simplex	H2	BPP-2015	↓
	chr04_2.23_s502	chr04	16	3.0	11.8	Simplex	H2	MRC-2015	↓
	chr04_70.05_s12931	chr04	139	3.2	13.7	Simplex	H2	MRC-2014	↑
	chr07_1.77_s639	chr07	12	3.8	17.4	Simplex	H2	MRC-2014	↓
	chr07_1.77_s639	chr07	15	3.1	13.1	Simplex	H2	MRC-2015	↓
	chr07_52.52_s10951	chr07	76	2.8	9.5	Simplex	H1	MRC-2014	↓
		chr07	76	4.3	19.3	Full Model		MRC-2015	
	chr10_50.67_d1945	chr10	60	2.6	8.6	Duplex	F14/V14	MRC-2014	↑
	chr10_44.85_s7410	chr10	45	2.7	8.3	Simplex	H1	BPP-2015	↑
TS	chr04_1.5_s247	chr04	10	4.0	17.6	Simplex	H2	MRC-2014	↓
	chr04_1.5_s247	chr04	12	3.1	10.4	Simplex	H2	BPP-2015	↓
	chr04_2.23_s502	chr04	16	4.5	21.7	Simplex	H2	MRC-2015	↓
	chr04_70.05_s12931	chr04	143	2.8	10.2	Full Model*		MRC-2014	↑
	chr07_1.77_s639	chr07	12	2.7	9.8	Simplex	H2	MRC-2015	↓
	chr07_52.52_s10951	chr07	76	2.7	9.6	Simplex	H1	MRC-2015	↓
	chr10_46.48_s7781	chr10	57	2.3	4.7	Simplex	H1	MRC-2014	↑
		chr10	60	2.3	6.7	Full Model		MRC-2015	
ATW	chr02_47.91_11358	chr02	105	2.8	8.4	Simplex	H4	MRC-2014	↓
	chr02_47.47_d1954	chr02	93	3.1	11.7	Duplex, dominant	D23	MRC-2015	↑
	chr04_70.05_s12931	chr04	143	2.5	8.5	Simplex	H2	MRC-2014	↑
	chr04_1.5_s247	chr04	10	2.6	7.3	Simplex	H2	BPP-2015	↓
	chr07_1.77_s639	chr07	12	3.9	16.9	Simplex	H2	MRC-2014	↓
		chr07	76	2.6	8.7	Full Model		MRC-2014	
	chr10_50.49_d1897	chr10	65	2.2	6	Duplex, additive	V14	MRC-2014	↑
	chr10_44.85_s7410	chr10	45	3.9	14.4	Simplex	H1	BPP-2015	↑
Vigor	chr02_47.91_11358	chr02	105	4.2	14.5	Simplex	H4	MRC-2014	↓
	chr02_46.24_10991	chr02	96	2.6	6.4	Simplex	H4	BPP-2015	↓
	chr04_1.5_s247	chr04	10	4.8	22.3	Simplex	H2	MRC-2014	↓
	chr04_0.02_d8	chr04	10	2.7	7.8	Duplex, additive	V13	BPP-2015	↑
	chr04_1.5_s247	chr04	10	3.4	14.6	Simplex	H2	MRC-2015	↓
	chr07_1.77_s639	chr07	14	2.9	11.5	Simplex	H2	MRC-2015	↓
		chr07	76	2.6	8.6	Duplex, no additive	F14	MRC-2014	
		chr07	76	2.9	11.1	Duplex, no additive	F14	MRC-2015	
	chr10_44.85_s7410	chr10	45	2.7	8.2	Simplex	H1	BPP-2015	↑
Height	chr02_47.91_11358	chr02	106	3.4	10.4	Simplex	H4	MRC-2014	↓
	chr02_46.24_10991	chr02	96	3.0	8.3	Simplex	H4	BPP-2015	↓
	chr04_1.5_s247	chr04	10	4.6	20.9	Simplex	H2	MRC-2014	↓
	chr04_0.02_d8	chr04	10	2.7	7.4	Duplex, additive	V13	BPP-2015	↑
	chr04_2.23_s502	chr04	16	3.3	13.4	Simplex	H2	MRC-2015	↓
	chr04_70.05_s12931	chr04	139	2.3	6.8	Simplex	H2	MRC-2014	↑
	chr07_1.77_s639	chr07	14	2.6	8.5	Simplex	H2	MRC-2015	↓
		chr07	76	2.5	7.8	Duplex, no additive	F14	MRC-2014	
		chr07	76	2.3	6.7	Duplex, no additive	F14	MRC-2015	
	chr10_44.85_s7410	chr10	47	2.6	9.2	Simplex	H1	MRC-2014	↑
	chr10_44.85_s7410	chr10	45	2.3	5.7	Simplex	H1	BPP-2015	↑

(Continued)

TABLE 7 | Continued

Trait	QTL locus	Chr	cM	LOD	% R2	QTL genetic model	Homologous chromosome	Site-year location	Main Q effect
Infl/plant	chr04_1.91_s379	chr04	10	3.5	13.3	Simplex	H3	BPP-2015	↑
	chr04_1.91_s379	chr04	10	2.6	8.5	Simplex	H3	MRC-2015	↑
Shape		chr06	93	2.5	7.9			MRC-2014	
		chr06	93	2.4	7.6	Duplex, no additive	F13	MRC-2015	
		chr10	60	5.1	18.6	Duplex, no additive	F13	MRC-2014	
		chr10	64	3.1	8.3	Duplex, no additive	F13	BPP-2015	
		chr10	60	4.3	15.5	Duplex, no additive		MRC-2015	
	chr11_42.28_s7725	chr11	142	3.5	14.8	Simplex	H2	MRC-2014	↓
	chr11_43.16_d2698	chr11	130	2.8	7.8	Duplex, no additive	F12	BPP-2015	↓
	chr11_42.28_s7725	chr11	135	4.4	21.5	Simplex	H2	MRC-2015	↓

*Total tuber yield (TTY) in g/plant, average tuber weight (ATW) in g, tuber set (TS) as number of tubers per plant, plant height (Height) in cm, plant vigor (Vigor) 1: low vigor, 5: high vigor, number of inflorescences per plant (Infl/plant), specific gravity (SPGR), and tuber shape (Shape) 1 =compressed, 2 =round, 3 = oval, 4 =oblong and 5 = long, main Q allele effect associated with lower (↓) or greater (↑) mean trait values. *Significant marker even though a specific model was not detected.*

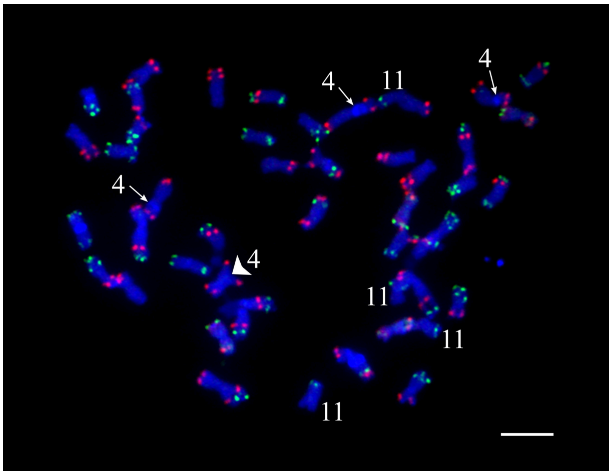


FIGURE 2 | Fluorescent *in situ* hybridization visualization of cv. Superior. chromosomes. The knobs on three copies of chromosome 4 are indicated by an arrow, the knob-less copy of chromosome 4 is indicated by a large arrowhead. Four copies of chromosome 11 are also indicated.

double reduction has occurred. In contrast, for the QTL on chromosome 10, the Q alleles in heterozygous genotypes were associated with greater mean values of these traits. Based on the analysis of the statistics of distribution of phenotypic data, dominance, intra-locus interactions, and epistatic interaction effects were considered as the main types of gene action associated with TTY and ATW, while a combination of additive, dominance, intra-locus interactions, and epistatic interaction effects was evident for TS, Height and Vigor. Either additive or dominant effects could explain the QTL with simplex allelic effects detected for most of the traits, while the duplex QTL effects were explained by dominant, additive and interaction effects.

We did not find any specific QTL for TTY and ATW, the traits with the greatest inbreeding depression. We hypothesize that probably multiple loci with a low percentage of explained variance as well as their epistatic interactions could be the reason underlying a lack of power to detect these QTL. Similarly, major QTL may not be segregating in this specific population. A clear example is the maturity locus on chromosome 5 associated with *Dof Zinc Finger Protein-StCDF* gene (Kloosterman et al., 2013). We did not identify a QTL in that region even though three alleles for cv. Superior were reported by Hardigan et al. (2017). The "Superior" alleles have polymorphisms (non-synonymous SNPs and truncations) compared to the allele associated with short day tuberization photoperiod control *CDF* in *Solanum tuberosum* Group Andigena. Therefore, all of these alleles should have similar additive effects in which any combination of those alleles in the diploid progeny is not associated with a segregating phenotype. Infl/plant corresponded to a trait for which we observed no inbreeding depression. The statistics of distribution analysis suggested that this trait should have gene actions associated with additive effects. Simplex and duplex with no additive allelic effects were the main type of gene action identified in the QTL analysis. Considering that several loci contribute to the genetic structure of a quantitative trait, we expect that epistatic interactions may play a major role in the genetic structure of the evaluated traits (best allelic combinations at different loci). In fact, only a few individuals in the progeny reached a genetic structure that generated a phenotype similar to the tetraploid parent.

The common QTLs for TTY, ATW, TS, Vigor and Height on chromosomes 2, 4, 7, and 10 co-localized with previous QTL reported for one or a few of the evaluated agronomic traits. Interestingly, the single parent tetraploid segregation revealed that the QTL were collectively associated with all of these traits. A QTL on chromosome 2 was reported for TTY and tuberization (Van den Berg et al., 1996; McCord et al., 2011; Manrique-Carpintero et al., 2015), on chromosome 4 for ATW, tuber size and tuberization (Van den Berg et al., 1996; D'hoop et al., 2014; Manrique-Carpintero et al., 2015), on chromosome 7 for tuber yield (Schäfer-Pregl et al., 1998), and on 10 for

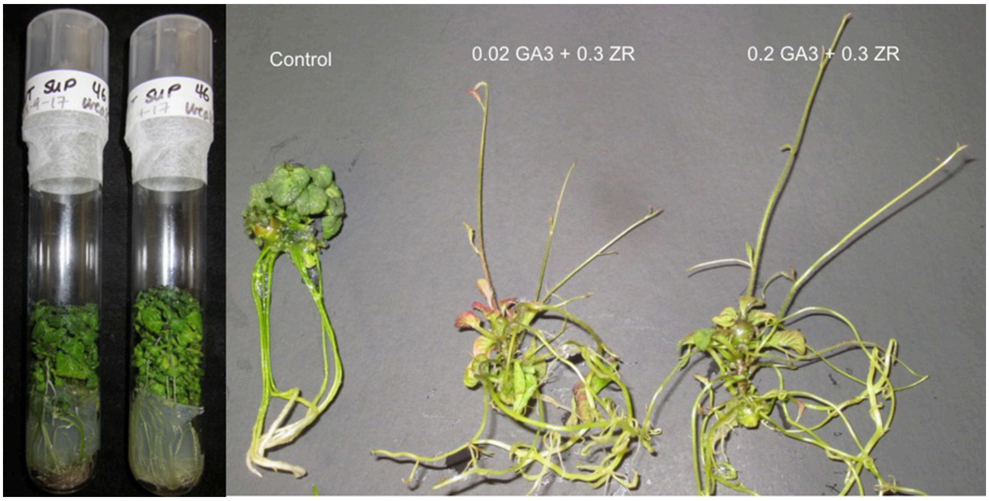


FIGURE 3 | Gibberellic acid treatment recovers a normal phenotype in the dwarf dihaploid VT_SUP_46. Comparison of plant growth in regular propagation medium (control), with medium supplemented with (0.02 mg/l gibberellic acid-GA$_3$ and 0.3 mg/l of zeatin riboside-ZR or 0.2 mg/l GA$_3$ and 0.3 mg/l ZR).

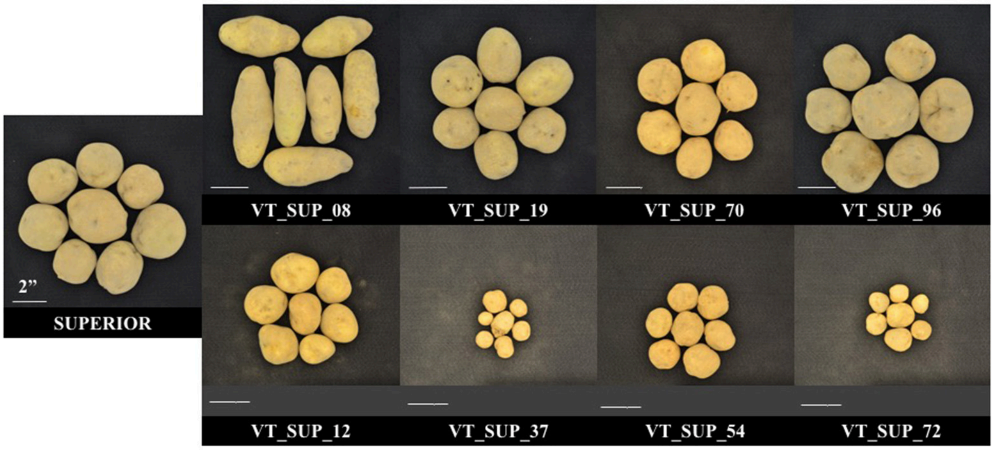


FIGURE 4 | Phenotypic differences in tuber size and shape between cv. Superior. and its dihaploid progeny with high **(Top)** and low agronomic performance **(Bottom)** under field season at Montcalm Research Center in 2015.

tuber yield, tuber set, and Vigor (Schäfer-Pregl et al., 1998; Manrique-Carpintero et al., 2015; Rak et al., 2017). Bonierbale et al. (1993) reported QTL on chromosomes 2, 4, and 7 for TTY, TS, and ATW, although the QTL on chromosome 2 does not match the chromosome arm location of our QTL. Similarly, several authors have reported a QTL on chromosome 10 for tuber shape (Van Eck et al., 1994; Prashar et al., 2014; Lindqvist-Kreuze et al., 2015), separating mainly compressed, round, and oval from the more elongated shape types oblong and long.

Importance of Inbreeding

Loss of heterozygosity has been associated with lower fitness. Considering that the homozygous alleles in the "Superior" parent were also homozygous in the progeny, we tested if the amount of segregating heterozygosity inherited from the tetraploid parent was associated with any trait. There was no correlation for most of the traits and poor correlation between the percentage of inherited heterozygosity and increasing TTY and TS trait values for all 3 site-years ($R^2 = 0.07$–0.09 and P-value < 0.03). As reported by Bonierbale et al. (1993), the additivity of a certain number of heterozygous loci rather than total heterozygosity makes a greater contribution to overall trait performance, along with the dominant alleles and epistatic effects. For instance, the weakest dihaploid clone (VT_SUP_46) had greater inherited heterozygosity than a high vigor and high-yielding dihaploid (VT_SUP_19), 60 and 55%, respectively. Haplotype analysis of "Superior" chromosomes showed a high level of heterogeneity in the parental genome. Cross-pollinated mating type, vegetative propagation, and polyploidy of cultivated potato contribute to retention of greater mutational load that is further complicated by rampant

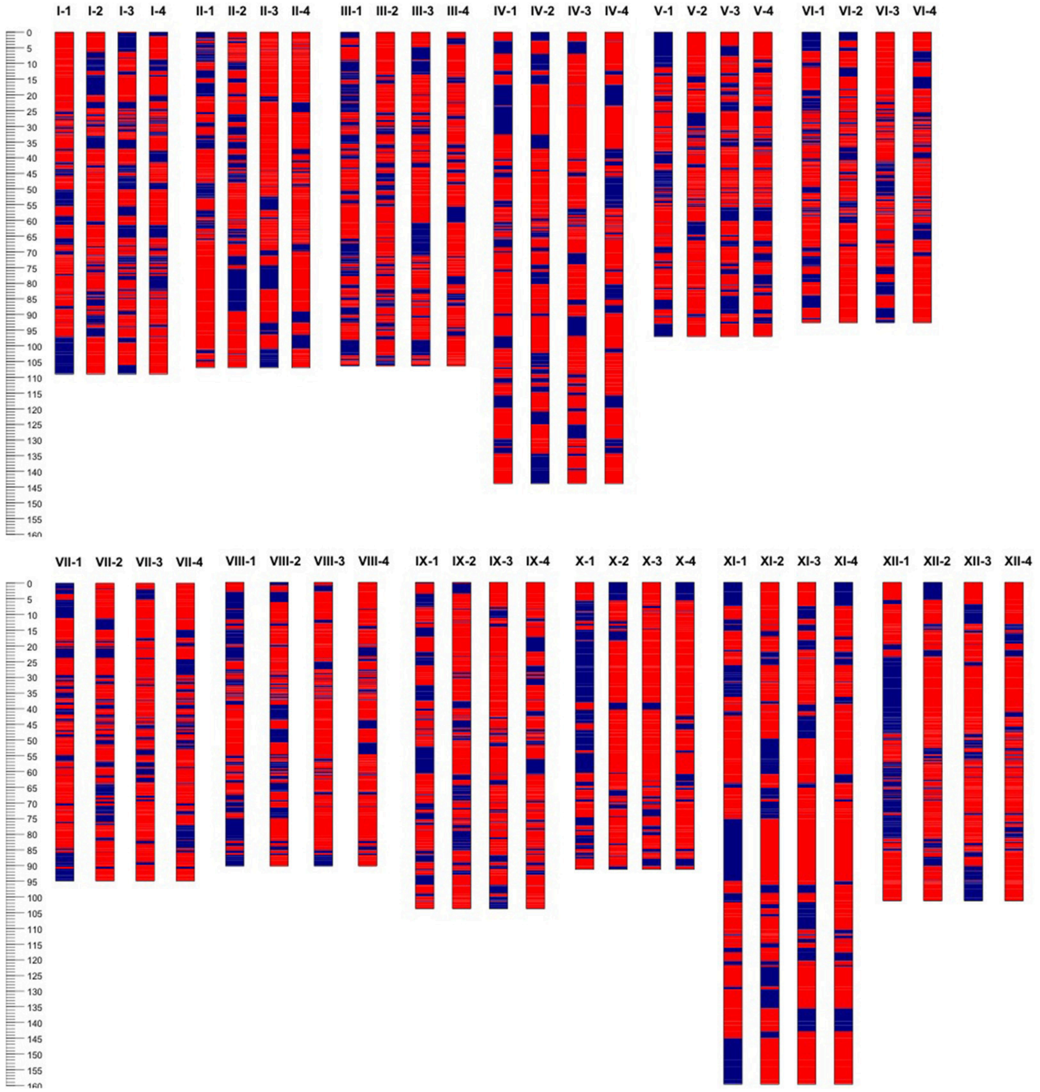


FIGURE 5 | Haplotype configuration of four homologous chromosomes (1–4) or each of 12 (I–XII) chromosomes of cv. Superior. using unique segregating single nucleotide polymorphic (SNP) bi-allelic (A, B, red and blue respectively) markers.

structural variation throughout the genome (Pham et al., 2017). Genetic load due to deleterious allelic mutations in the simplex configuration could be compensated by the alternative allele, but also by multiple loci with similar function(s) in the polyploid genome. At a given locus, it is possible to have: (i) duplicate alleles or alleles with synonymous nucleotide polymorphisms that will not affect the functionality at the protein level, (ii) alleles with polymorphisms that alter functionality at the protein level, and /or (iii) alleles with no functionality (i.e., a null allele). In principle, any alternative functional allele would compensate for dysfunctionality in a dihaploid or tetraploid individual when present with the lethal allele, at the same or different locus. Therefore, the combination of alleles at multiple loci determines the trait phenotype. However, epistasis complicates the identification of associations between markers and phenotypic performance (Ceballos et al., 2015). Inbreeding can be the most efficient method to organize the genome to combine favorable alleles interacting in a stable epistatic system, therefore high fitness progeny would have the best genetic structure (Jansky et al., 2016). By design, we examined only biallelic SNPs, thereby disregarding the contributions of multiallelic loci on yield attributing traits. For triallelic loci, 1/6th of the dihaploid progeny would be expected to be homozygous at any given SNP site whereas for tetraallelic loci, all dihaploid progeny would remain heterozygous, albeit with different combinations of alleles.

Candidate Genes

In the common QTL regions identified in this study (**Figure 1**) for TTY, ATW, TS, Height and Vigor, we hypothesized that candidate genes associated with overall plant growth and

TABLE 8 | Genetic distance between homologs (H) of each chromosome (Chr) of cv. Superior.

Homologous pairwise comparison	Chr01	Chr02	Chr03	Chr04	Chr05	Chr06	Chr07	Chr08	Chr09	Chr10	Chr11	Chr12
H1–H2	0.58	0.76	0.73	0.61	0.67	0.59	0.67	0.7	0.62	0.81	0.74	0.77
H1–H3	0.65	0.55	0.58	0.55	0.67	0.58	0.51	0.62	0.51	0.81	0.54	0.77
H1–H4	0.6	0.55	0.57	0.45	0.66	0.52	0.65	0.51	0.57	0.72	0.58	0.6
H2–H3	0.57	0.47	0.47	0.52	0.52	0.61	0.57	0.54	0.58	0.3	0.53	0.39
H2–H4	0.62	0.48	0.41	0.61	0.48	0.46	0.42	0.53	0.49	0.26	0.45	0.34
H3–H4	0.37	0.3	0.49	0.53	0.23	0.48	0.44	0.42	0.46	0.24	0.41	0.31

Distance color code: shorter genetic distance toward red while greater genetic distance greener.

development, as well as tuberization (Supplementary Table 3) would be present. Hormonal regulation, sucrose metabolism, photoperiod, circadian clock, and age-dependent signaling pathways are involved in tuber initiation and growth (Navarro et al., 2015) for which some genes have been identified. In the QTL region on chromosomes 2 and 7, candidate genes in the photoperiod regulatory pathway associated with length of plant cycle and tuberization were identified (*Dof Zinc Finger Protein-StCDF3*, *CONSTANTS-CO*, and mi*RNA156*) around 46 and 2 Mb, respectively. High accumulation of sucrose and starch in terminal sink organs is enhanced by efflux from the leaves promoting tuberization, down-regulation of the phloem *Sucrose transporter 4 (SUT4)* gene is critical to the switch from apoplastic to symplastic phloem uploading (Chincinska et al., 2013). *SUT4* follows a circadian expression pattern, has reciprocal regulation with gibberellic acid (GA), and affects the expression of circadian-regulated genes, flowering, tuberization and shade avoidance. *SUT4* is located at 65.8 Mb on chromosome 4, in the region where a QTL was detected. The breakdown of active GA is required for tuberization and gibberellin 2-oxidase genes are part of the mechanism that controls endogenous levels of GA (Kloosterman et al., 2007); we identified a *Gibberellin 2-oxidase 2 (GA2ox2)* candidate gene at 51.9 Mb in a QTL regions on chromosome 7. Interestingly, in the other QTL region of chromosome 7 at 1.9 Mb is a *Trehalose-phosphate synthase 1* (TPS1) gene with a potential role in the T6P regulatory pathway that was recently associated with flowering and tuberization in potato (Seibert et al., 2017). Ectopic expression of *Lonely Guy 1* (*LOG1*), a cytokinin-activating enzyme, drove the formation of aerial minitubers in tomato (Eviatar-Ribak et al., 2013). The plants displayed a unique transcriptome signaling network probably associated with the appropriated local hormonal balance for tuber formation. Differential expression and pleiotropic effects of *LOG* genes showed their major role in cytokinin metabolism to modulate plant growth and development in *Arabidopsis thaliana* (Kuroha et al., 2009). A cytokinin riboside 5′-monophosphate phosphoribohydrolase *LOG3* gene is located in the QTL region at 56 Mb on chromosome 10. For tuber shape, several candidate genes associated with cell structure and function, and pectin metabolism have been reported in the major QTL located around 48 Mb on chromosome 10 (Lindqvist-Kreuze et al., 2015). Similarly, in the QTL region discovered in our analysis on chromosome 6, a *Pectinesterase* gene is located at 58 Mb.

Dwarf Phenotype

There is strong evidence that a dwarf phenotype observed in our "Superior" dihaploid population is the result of GA_3 deficiency. The dark green and rosette dwarf phenotype has been reported in potato in hybrid progeny of cv. Superior. as well as some other potato diploid and tetraploid clones (Bamberg and Hanneman, 1991; Valkonen et al., 1999). In all cases, reversion of the dwarf phenotype occurred following GA_3 application. A single recessive locus encoding *ga1* was proposed to cause the dwarf phenotype, which was confirmed by evaluation of test segregation in several crosses (Bamberg and Miller, 2012). The study also revealed that a gibberellin deficiency allele was in simplex configuration (GGGg) in "Superior." The homozygous state gg of the recessive allele of a simplex locus in a dihaploid population is expected only due to double reduction, therefore a small proportion of dwarf phenotype would be observed in the dihaploid progeny. In fact, VT_SUP_46 is a unique clone in our "Superior" dihaploid population with a strong dwarf phenotype. Examination of the regions with potential double reduction in VT_SUP_46, revealed the end of chromosomes 6 as the candidate region. However, a few other clones also showed double reduction but did not have dwarf phenotype, suggesting that other loci could compensate the GA_3 supply in those dihaploid clones. When *in vitro* plantlets of VT_SUP_46 were grown on propagation medium supplemented with GA (0.02 and 0.2 mg/l) the plants elongated to a normal phenotype (**Figure 3**).

CONCLUSION

Genetic load in the "Superior" cultivar was unmasked through the generation of a dihaploid population. The segregation of the parental tetraploid configuration identified major QTL regions associated with most of the evaluated agronomic traits. Interestingly, four chromosomes were identified with common QTL that could elucidate interconnected metabolism. Candidate genes regulating plant development and tuberization were identified in the QTL regions. Complementation of gene function due to homozygous deleterious alleles could play a major role in trait performance in polyploid potato.

AUTHOR CONTRIBUTIONS

RV, CRB, and DD planned and designed the project. NM-C drafted the manuscript. JC and NM-C conducted phenotypic data analysis, linkage mapping, and QTL analysis. NM-C tissue culture experiment. GB and JJ made cytogenetic analysis. DD and JC were involved field experiments. GP generate genotypic data. RV and FL generated the Superior dihaploid population. All authors contributed to the editing of the manuscript and approved the final draft.

FUNDING

This research was supported by the National Science Foundation under Grant No. IOS-1237969 to CRB, DD, Yuehua Cui, and RV.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00944/full#supplementary-material

REFERENCES

Bamberg, J. B., and Hanneman, R. E. (1991). Characterization of a new gibberellin related dwarfing locus in potato (*Solanum tuberosum* L.). *Am. Potato J.* 68, 45–52. doi: 10.1007/BF02893340

Bamberg, J., and Miller, J. C. (2012). Comparisons of *ga1* with other reputed gibberellin mutants in potato. *Am. J. Potato Res.* 89, 142–149. doi: 10.1007/s12230-012-9236-5

Bonierbale, M. W., Plaisted, R. L., and Tanksley, S. D. (1993). A test of the maximum heterozygosity hypothesis using molecular markers in tetraploid potatoes. *Theor. Appl. Genet.* 86, 481–491. doi: 10.1007/BF00838564

Braz, G. T., He, L., Zhao, H. N., Zhang, T., Semrau, K., Rouillard, J. M., et al. (2018). Comparative oligo-FISH mapping: an efficient and powerful methodology to reveal karyotypic and chromosomal evolution. *Genetics* 208, 513–523. doi: 10.1534/genetics.117.300344

Carputo, D., and Frusciante, L. (2011). "Classical genetics and traditional breeding," in *Genetics, Genomics and Breeding of Potato,* eds J.M. Bradeen and C. Kole (Enfield: Science Publishers, Inc), 20-40.

Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., and Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *Theor. Appl. Genet.* 128, 1647–1667. doi: 10.1007/s00122-015-2555-4

Chakravarti, A. (1991). A graphical representation of genetic and physical maps - the Marey map. *Genomics* 11, 219–222. doi: 10.1016/0888-7543(91)90123-V

Chase, S. S. (1963). Analytic breeding in *Solanum tuberosum* L - a scheme utilizing parthenotes and other diploid stocks. *Can. J. Genet. Cytol.* 5, 359–363. doi: 10.1139/g63-049

Chincinska, I., Gier, K., Krugel, U., Liesche, J., He, H. X., Grimm, B., et al. (2013). Photoperiodic regulation of the sucrose transporter StSUT4 affects the expression of circadian-regulated genes and ethylene production. *Front. Plant Sci.* 4:26. doi: 10.3389/Fpls.2013.00026

Cipar, M. S., and Lawrence, C. H. (1972). Scab resistance of haploids from two *Solanum tuberosum* cultivars. *Am. Potato J.* 49, 117–119. doi: 10.1007/BF02868225

Da Silva, W. L., Ingram, J., Hackett, C. A., Coombs, J. J., Douches, D., Bryan, G. J., et al. (2017). Mapping loci that control tuber and foliar ssymptoms caused by PVY in autotetraploid potato (*Solanum tuberosum* L.). *G3 (Bethesda)* 7, 3587–3595. doi: 10.1534/g3.117.300264

De Jong, H., and Rowe, P. R. (1971). Inbreeding in cultivated diploid potatoes. *Potato Res.* 14, 74–83. doi: 10.1007/BF02355931

De Maine, M. J. (1984). Patterns of variation in potato dihaploid families. *Potato Res.* 27, 1–11. doi: 10.1007/BF02356192

D'hoop, B. B., Keizer, P. L. C., Paulo, M. J., Visser, R. G. F., Van Eeuwijk, F. A., and Van Eck, H. J. (2014). Identification of agronomically important QTL in tetraploid potato cultivars using a marker-trait association analysis. *Theor. Appl. Genet.* 127, 731–748. doi: 10.1007/s00122-013-2254-y

Eviatar-Ribak, T., Shalit-Kaneh, A., Chappell-Maor, L., Amsellem, Z., Eshed, Y., and Lifschitz, E. (2013). A cytokinin-activating enzyme promotes tuber formation in tomato. *Curr. Biol.* 23, 1057–1064. doi: 10.1016/j.cub.2013.04.061

Fasoulas, A. C. (1988). *The Honeycomb Methodology of Plant Breeding.* Thessaloniki: A. C. Fasoulas.

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., et al. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE* 7:e36347. doi: 10.1371/journal.pone.0036347

Golmirzaie, A. M., Bretschneider, K., and Ortiz, R. (1998). Inbreeding and true seed in tetrasomic potato. II. Selfing and sib-mating in heterogeneous hybrid populations of *Solanum tuberosum. Theor. Appl. Genet.* 97, 1129–1132. doi: 10.1007/s001220051001

Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F., and Milne, I. (2017). TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *J. Hered.* 108, 438–442. doi: 10.1093/jhered/esx022

Hackett, C. A., McLean, K., and Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS ONE* 8:e63939. doi: 10.1371/journal.pone.0063939

Hardigan, M. A., Crisovan, E., Hamilton, J. P., Kim, J., Laimbeer, P., Leisner, C. P., et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell* 28, 388–405. doi: 10.1105/tpc.15.00538

Hardigan, M. A., Laimbeer, F. P. E., Newton, L., Crisovan, E., Hamilton, J. P., Vaillancourt, B., et al. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9999–E10008. doi: 10.1073/pnas.1714380114

Hermsen, J. G. T., and Verdenius, J. (1973). Selection from *Solanum tuberosum* Group Phureja of genotypes combining high frequency haploid Induction with homozygosity for embryo spot. *Euphytica* 22, 244–259. doi: 10.1007/BF00022632

Hutten, R. C. B., Soppe, W. J. J., Hermsen, J. G. T., and Jacobsen, E. (1995). Evaluation of dihaploid populations from potato varieties and breeding lines. *Potato Res.* 38, 77–86. doi: 10.1007/BF02358072

Jansky, S. H., Charkowski, A. O., Douches, D. S., Gusmini, G., Richael, C., Bethke, P. C., et al. (2016). Reinventing potato as a diploid inbred line-based crop. *Crop Sci.* 56, 1412–1422. doi: 10.2135/cropsci2015.12.0740

Kloosterman, B., Abelenda, J. A., Gomez, M. D. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* 495, 246–250. doi: 10.1038/Nature11912

Kloosterman, B., Navarro, C., Bijsterbosch, G., Lange, T., Prat, S., Visser, R. G. F., et al. (2007). StGA2ox1 is induced prior to stolon swelling and controls GA levels during potato tuber development. *Plant J.* 52, 362–373. doi: 10.1111/j.1365-313X.2007.03245.x

Kotch, G. P., Ortiz, R., and Peloquin, S. J. (1992). Genetic analysis by use of potato haploid populations. *Genome* 35, 103–108. doi: 10.1139/g92-018

Kuroha, T., Tokunaga, H., Kojima, M., Ueda, N., Ishida, T., Nagawa, S., et al. (2009). Functional analyses of LONELY GUY cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. *Plant Cell* 21, 3152–3169. doi: 10.1105/tpc.109.068676

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [Preprint] arXiv:1303.3997v1 [q-bio.GN].*

Lindqvist-Kreuze, H., Khan, A., Salas, E., Meiyalaghan, S., Thomson, S., Gomez, R., et al. (2015). Tuber shape and eye depth variation in a diploid family of Andean potatoes. *BMC Genet.* 16:57. doi: 10.1186/s12863-015-0213-0

Manrique-Carpintero, N. C., Coombs, J. J., Cui, Y., Veilleux, R. E., Buell, C. R., and Douches, D. (2015). Genetic map and quantitative trait locus analysis of agronomic traits in a diploid potato population using single nucleotide polymorphism markers. *Crop Sci.* 55, 2566–2579. doi: 10.2135/cropsci2014.10.0745

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Massa, A. N., Manrique-Carpintero, N. C., Coombs, J. J., Zarka, D. G., Boone, A. E., Kirk, W. W., et al. (2015). Genetic linkage mapping of economically important traits in cultivated tetraploid potato (*Solanum tuberosum* L.). *G3 (Bethesda)* 5, 2357–2364. doi: 10.1534/g3.115.019646

Matsubayashi, M. (1979). Genetic variation in diploid potato clones, with special reference to phenotypic segregations in some characters. *Sci. Rept. Fac. Agr. Kobe Univ.* 13, 185–172.

McCord, P. H., Sosinski, B. R., Haynes, K. G., Clough, M. E., and Yencho, G. C. (2011). QTL mapping of internal heat necrosis in tetraploid potato. *Theor. Appl. Genet.* 122, 129–142. doi: 10.1007/s00122-010-1429-z

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

Mendoza, H. A., and Haynes, F. L. (1974). Genetic basis of heterosis for yield in the autotetraploid potato. *Theor. Appl. Genet.* 45, 21–25. doi: 10.1007/BF00281169

Miranda Filho, J. B. (1999). "Inbreeding and heterosis," in *Genetics and Exploitation of Heterosis in Crops,* eds J. G. Coors and S. Pandey (Madison, WI: American Society of Agronomy).

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* 15, 473–497. doi: 10.1111/j.1399-3054.1962.tb08052.x

Navarro, C., Cruz-Oro, E., and Prat, S. (2015). Conserved function of flowering locus T (FT) homologues as signals for storage organ differentiation. *Curr. Opin. Plant Biol.* 23, 45–53. doi: 10.1016/j.pbi.2014.10.008

Owen, H. R., Veilleux, R. E., Levy, D., and Ochs, D. L. (1988). Environmental, genotypic, and ploidy effects on endopolyploidization within a genotype of *Solanum phureja* and its derivatives. *Genome* 30, 506–510. doi: 10.1139/g88-085

Peloquin, S. J., and Hougas, R. W. (1960). Genetic variation among haploids of the common potato. *Am. Potato J.* 37, 289–297. doi: 10.1007/BF02855072

Peloquin, S. J., Werner, J. E., and Yerk, G. L. (1991). "The use of potato haploids in genetics and breeding," in *Chromosome Engineering in Plants: Genetics, Breeding, Evolution,* Vol. 2, Part B, eds T. Tsuchiya and P. K. Gupta (Amsterdam: Elsevier Science Publishers B.V.), 79–92.

Pham, G. M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D. S., and Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J.* 92, 624–637. doi: 10.1111/tpj.13706

Pineda, O., Bonierbale, M., and Plaisted, R. (1993). Identification of RFLP markers linked to the H1 gene conferring resistance to the potato cyst nematode *Globodera rostochiensis*. *Genome* 36, 152–156. doi: 10.1139/g93-019

Prashar, A., Hornyik, C., Young, V., McLean, K., Sharma, S. K., Dale, M. F. B., et al. (2014). Construction of a dense SNP map of a highly heterozygous diploid potato population and QTL analysis of tuber shape and eye depth. *Theor. Appl. Genet.* 127, 2159–2171. doi: 10.1007/s00122-014-2369-9

Preedy, K. F., and Hackett, C. A. (2016). A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theor. Appl. Genet.* 129, 2117–2132. doi: 10.1007/s00122-016-2761-8

Rak, K., Bethke, P. C., and Palta, J. P. (2017). QTL mapping of potato chip color and tuber traits within an autotetraploid family. *Mol. Breed.* 37:15. doi: 10.1007/s11032-017-0619-7

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi,-, C., Bredeson, J. V., et al. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* 49, 959–963. doi: 10.1038/ng.3845

Rieman, G. H. (1962). Superior: a new white, medium-maturiting, scab-resistant potato variety with high chipping quality. *Am. Potato J.* 39, 19–28. doi: 10.1007/BF02912628

Rokka, V. M. (2009). "Potato haploids and breeding," in *Advances in Haploid Production in Higher Plants,* eds A. Touraev, B.P. Forster and S.M. Jain (Dordrecht: Springer), 199–208.

Schäfer-Pregl, R., Ritter, E., Concilio, L., Hesselbach, J., Lovatti, L., Walkemeier, B., et al. (1998). Analysis of quantitative trait loci (QTLs) and quantitative trait alleles (QTAs) for potato tuber yield and starch content. *Theor. Appl. Genet.* 97, 834–846. doi: 10.1007/s001220050963

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Seibert, T., Abel, C., Feil, R., Prat, S., and Wahl, V. (2017). "The role of the T6P pathway during flowering and tuberization in *Solanum tuberosum*," in *XIV Solanaceae and 3rd Cucurbitaceae Joint Conference* (Valencia).

Sharma, S. K., Bolser, D., de Boer, J., Sonderkaer, M., Amoros, W., Carboni, M. F., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3* 3, 2031–2047. doi: 10.1534/g3.113.007153

Song, Y. S., Hepting, L., Schweizer, G., Hartl, L., Wenzel, G., and Schwarzfischer, A. (2005). Mapping of extreme resistance to PVY (Ry(sto)) on chromosome XII using anther-culture-derived primary dihaploid potato lines. *Theor. Appl. Genet.* 111, 879–887. doi: 10.1007/s00122-005-0010-7

Valkonen, J. P. T., Moritz, T., Watanabe, K. N., and Rokka, V. M. (1999). Dwarf (di)haploid *pito* mutants obtained from a tetraploid potato cultivar (*Solanum tuberosum* subsp *tuberosum*) via anther culture are defective in gibberellin biosynthesis. *Plant Sci.* 149, 51–57. doi: 10.1016/S0168-9452(99)00141-7

Van Berloo, R. (2008). GGT 2.0: versatile software for visualization and analysis of genetic data. *J. Hered.* 99, 232–236. doi: 10.1093/jhered/esm109

Van den Berg, J. H., Ewing, E. E., Plaisted, R. L., McMurry, S., and Bonierbale, M. W. (1996). QTL analysis of potato tuberization. *Theor. Appl. Genet.* 93, 307–316. doi: 10.1007/BF00223170

Van Eck, H. J., Jacobs, J. M. E., Stam, P., Ton, J., Stiekema, W. J., and Jacobsen, E. (1994). Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. *Genetics* 137, 303–309.

Van Ooijen, J. W. (2006). *JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations.*

Varga, M. B., and Ferenczy, L. (1956). Effect of rindite on the development of the growth-substances in potato tubers. *Nature* 178, 1075–1075. doi: 10.1038/1781075a0

Velasquez, A. C., Mihovilovich, E., and Bonierbale, M. (2007). Genetic characterization and mapping of major gene resistance to potato leafroll virus in *Solanum tuberosum* ssp *andigena*. *Theor. Appl. Genet.* 114, 1051–1058. doi: 10.1007/s00122-006-0498-5

Wang, L., Beissinger, T. M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., and Hufford, M. B. (2017). The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* 18:215. doi: 10.1186/s13059-017-1346-4

Conflict of Interest Statement: 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.

Decomposing Additive Genetic Variance Revealed Novel Insights into Trait Evolution in Synthetic Hexaploid Wheat

Abdulqader Jighly [1,2]*, ***Reem Joukhadar*** [1,3], ***Sukhwinder Singh*** [4] **and** ***Francis C. Ogbonnaya*** [5]

[1] Agriculture Victoria, Agriculture Research Division, AgriBio, Centre for AgriBiosciences, Bundoora, VIC, Australia, [2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, Australia, [3] Department of Animal, Plant and Soil Sciences, La Trobe University, Bundoora, VIC, Australia, [4] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [5] Grains Research and Development Corporation, Kingston, ACT, Australia

OPEN ACCESS

Edited by:
Richard John Abbott,
University of St Andrews,
United Kingdom

Reviewed by:
Margarida Matos,
Universidade de Lisboa, Portugal
Zhihong Zhu,
The University of Queensland,
Australia

****Correspondence:***
Abdulqader Jighly
abdulqader.jighly@ecodev.vic.gov.au

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

Received: *11 September 2017*
Accepted: *22 January 2018*
Published: *06 February 2018*

Citation:
Jighly A, Joukhadar R, Singh S and Ogbonnaya FC (2018) Decomposing Additive Genetic Variance Revealed Novel Insights into Trait Evolution in Synthetic Hexaploid Wheat. Front. Genet. 9:27. doi: 10.3389/fgene.2018.00027

Whole genome duplication (WGD) is an evolutionary phenomenon, which causes significant changes to genomic structure and trait architecture. In recent years, a number of studies decomposed the additive genetic variance explained by different sets of variants. However, they investigated diploid populations only and none of the studies examined any polyploid organism. In this research, we extended the application of this approach to polyploids, to differentiate the additive variance explained by the three subgenomes and seven sets of homoeologous chromosomes in synthetic allohexaploid wheat (SHW) to gain a better understanding of trait evolution after WGD. Our SHW population was generated by crossing improved durum parents (*Triticum turgidum;* $2n = 4x = 28$, AABB subgenomes) with the progenitor species *Aegilops tauschii* (syn *Ae. squarrosa*, *T. tauschii*; $2n = 2x = 14$, DD subgenome). The population was phenotyped for 10 fungal/nematode resistance traits as well as two abiotic stresses. We showed that the wild D subgenome dominated the additive effect and this dominance affected the A more than the B subgenome. We provide evidence that this dominance was not inflated by population structure, relatedness among individuals or by longer linkage disequilibrium blocks observed in the D subgenome within the population used for this study. The cumulative size of the three homoeologs of the seven chromosomal groups showed a weak but significant positive correlation with their cumulative explained additive variance. Furthermore, an average of 69% for each chromosomal group's cumulative additive variance came from one homoeolog that had the highest explained variance within the group across all 12 traits. We hypothesize that structural and functional changes during diploidization may explain chromosomal group relations as allopolyploids keep balanced dosage for many genes. Our results contribute to a better understanding of trait evolution mechanisms in polyploidy, which will facilitate the effective utilization of wheat wild relatives in breeding.

Keywords: polyploidy, synthetic hexaploid wheat, diploidization, additive variance, heritability

INTRODUCTION

Polyploidization, whole genome duplication (WGD), is a natural process in which a single genome can be duplicated to form autopolyploids with more than two homologs for each chromosome, or multiple genomes are duplicated following hybridization between two or more species to form allopolyploids with multiple pairs of homologs derived from different ancestral genomes, termed homoeologs. Following WGD, multiple copies of duplicated genes may be lost, diverge in function, or silenced through a phenomenon called "diploidization" in which balanced dosages for many genes can be retrieved (Ohno, 1970; Lynch and Conery, 2000; Tate et al., 2009; Conant et al., 2014). Rapid genomic rearrangements and epigenetic changes have been observed directly after WGD (Ozkan et al., 2001; Shaked et al., 2001; Kashkush et al., 2002; Hegarty et al., 2008) which can cause changes in the architecture of different traits (Weiss-Schneeweiss et al., 2013).

WGD can be induced in laboratories to generate new taxa such as triticale (Stace, 1987), or to introduce new variation into known taxa such as bread wheat (*Triticum aestivum*, 2n = 6x = 42, AABBDD) which suffered a severe genetic bottleneck during its origin (Yang et al., 2009). Synthetic hexaploid wheat (SHW) can be generated by crossing *Triticum turgidum* (2n = 4x = 28, AABB) with *Aegilops tauschii* (2n = 2x = 14, DD), mimicking the natural evolutionary origin of bread wheat. SHW germplasm is a proven source of genetic diversity to improve yield (Gororo et al., 2002; Dreccer et al., 2007; Ogbonnaya et al., 2007, 2013), soil-borne pathogen (Mulki et al., 2013), insect (El-Bouhssini et al., 2013; Joukhadar et al., 2013), and fungal disease resistance (Zegeye et al., 2014; Jighly et al., 2016), as well as boron (Emebiri and Ogbonnaya, 2015) and salinity tolerance (Dreccer et al., 2004; Ogbonnaya et al., 2008a). However, it remains uncertain how the three subgenomes (A, B, and D) of bread wheat contribute to observed phenotypes or whether the wild *Aegilops* parent makes a considerable contribution to the additive genetic variance for different traits especially when crossed with an improved or elite durum wheat parent. This can be investigated by partitioning the total additive trait variance into different chromosomes in a SHW population.

Recently, a number of studies partitioned the additive variance of different traits captured by multiple sets of markers in both human and animal quantitative genetics studies. Applications varied from differentiating the variance captured by different chromosomes (Robinson et al., 2013), genotyped, and imputed variants (Lee et al., 2012), genic, and intergenic variants (Yang et al., 2011b), different SNP chips (Chen et al., 2014), to differentiating the variance of common and rare variants (Lee et al., 2013; Yang et al., 2015). In general, almost all studies reported a medium to high correlation between chromosome size and its explained additive variance for the studied traits. Yet, this approach has not been applied to any plant population, particularly among polyploid species such as wheat, where considerable efforts have gone into exploiting valuable sources of new genes from its progenitor species for cultivated wheat improvement (Ogbonnaya et al., 2013). Applying this approach to allopolyploids can provide a better understanding and a new way for differentiating the additive effects captured by different subgenomes.

In this research, we used a SHW population to investigate the contribution of each subgenome to trait variation. The SHW population was derived from crosses between wild *Ae. tauschii* parents and improved durum cultivars and was phenotyped for resistance to 10 different diseases and tolerance to two abiotic stresses. The same dataset was previously characterized in multiple genome-wide association studies (GWAS) for major genes associated with these different stresses (Mulki et al., 2013; Emebiri and Ogbonnaya, 2015; Jighly et al., 2016). However, the GWAS approach does not adequately provide the precise contribution of each chromosome/subgenome to the total heritability as genes identified through GWAS represent only a small proportion of the total heritability (Goldstein, 2009; Yang et al., 2017). Such information is critical to understanding trait evolution in newly synthesized allopolyploids and to efficiently utilize wild relatives in wheat breeding. In the present paper, we investigated this by partitioning the additive variance into each of the 21 SHW chromosomes. The relation between partitioned additive variance and chromosome, subgenome and chromosomal group size was also investigated. To the best of our knowledge, this is the first study to use this approach in polyploid or plant populations.

MATERIALS AND METHODS

SHW Phenotyping and Genotyping

The SHW population consists of 173 crosses between different *A. tauschii* accessions and elite durum cultivars (**Table S1**). The population was genotyped with DArTSeq—a genotyping by sequencing, (GBS) approach, developed by Diversity Array Technology, DArT, http://www.diversityarrays.com/. The full method is described in Sehgal et al. (2015). In brief, restriction enzymes were used first to reduce the complexity of the wheat genome and the *Pst1-RE* adapters were tagged with 96 barcodes. This strategy allows for multiplexing 96 samples in a single Illumina HiSeq2500 lane to generate around 0.5 million of 77 bp reads per sample. The generated FASTQ files were trimmed at Phred score 30 and further filtering steps and SNP calling were conducted using designed scripts developed by DArT P/L. Only SNPs with <20% missing data and >5% minor allele frequency were used in subsequent analyses. The SNP dataset used for the current study was previously published as a supplement in Jighly et al. (2016).

The SHW population was phenotyped for aluminum (Al) and boron (Br) tolerance, stem (Sr), yellow (Yr) and leaf (Lr) rusts, crown rot (Cr), yellow leaf spot (YLS), *septoria nodorum* leaf blotch (SNL) and *septoria nodorum* glume blotch (SNG), root lesion nematodes [*Pratylenchus neglectus* (Pn) and *Pratylenchus thornei* (Pt)] and cereal cyst nematode (CCN) resistance. Experimental details were previously described in (Ogbonnaya et al., 2008b; Emebiri and Ogbonnaya, 2015; Jighly et al., 2016). Briefly, the germplasm was screened in three replicates for the three rust diseases under field conditions. The most commercially important fungal pathotypes used for infection were 104–1,2,3,(6), (7), 11, 13 (accession number

200347) for Lr; 98–1,2,3,5,6 (accession number 781219) for Sr; and 134 E16A (021510) for Yr. Four different isolates (WAC 4302, WAC 4305, WAC 4306, and WAC 4309) were used in four replicates under greenhouse conditions for SNG and SNL. YLS was also screened in a controlled environment against isolates 03–0148, 03–0152, and 03–0053. For CCN, plants were considered resistant if they had less than five cysts per plant root while plants were considered susceptible if they had more than 30 cysts. Plants with 5–30 cysts were considered moderately resistant to moderately susceptible. The severity of Pn and the number of Pt nematodes per plant were used to infer the score of resistance by comparing the plant response to resistant and susceptible checks. Br tolerance was phenotyped by measuring root growth at the seedling stage on a filter paper soaked with boron while Al tolerance was measured using the hematoxylin staining of root apices method (Raman et al., 2010).

Statistical Analysis

We estimated 21 genetic relatedness matrices (GRMs) from SNPs located on each one of the SHW chromosomes following the method described in (Yang et al., 2010, 2011a). The variance explained by each chromosome was estimated using the genomic-relatedness-based restricted maximum likelihood (GREML) analysis by fitting all 21 GRMs simultaneously in the mixed linear model (Lee et al., 2012; Lee and van der Werf, 2016):

$$y = X\beta + \sum_{i=1}^{n} g_i + \varepsilon$$

Where y is a vector of phenotypes, n is the number of chromosomes (21 in our case), β is a vector of fixed effects, X is an incidence matrix that relates individuals to fixed effects and ε is a vector of random errors. g_i is a vector of random additive genetic effect attribute to chromosome i. The variance structure of phenotype is equal to:

$$V = \sum_{i=1}^{n} A_i \sigma_{g_i}^2 + I\sigma_e^2$$

Where A_i is the GRM for chromosome i, $\sigma_{g_i}^2$ is the additive genetic variance captured by SNPs on chromosome i, I is an identity matrix and σ_e^2 is the error variance.

We ran the analysis twice, with and without including the first 10 principal components (PCs) as fixed effects. Including a number of PCs in the model can control for population structure in the germplasm; thus, the effect of population structure will be minimal if the model that fits PCs revealed similar results to the model that does not include PCs (Lee et al., 2012). The first 10 PCs were calculated using PLINK 1.9 (http://www.cog-genomics.org/plink/1.9/). To further investigate the effect of the correlation between different chromosomes due to shared structure among chromosomes (Lee et al., 2012; Yang et al., 2017), we calculated the conditional effect for each one based on the other 20 chromosomes. This was done by fitting 21 different models that each excluded one different GRM from the joint analysis. If the SNPs located on the excluded chromosome were correlated with SNPs on the other 20 chromosomes, the conditional effect analysis will overestimate the additive variance for the 20 chromosomes. Subtracting the conditional additive variance from the overall additive variance inferred from the full model is equal to the proportion of additive variance of the excluded chromosome that is not correlated with other chromosomes. This value can be used to investigate dependency among chromosomes and to confirm differences among subgenomes.

The D subgenome in our germplasm had very large LD blocks compared to the A and B subgenomes (Jighly et al., 2016) which may overestimate the heritability for the D subgenome (Speed et al., 2012). Thus, we repeated the analysis after randomly omitting 20% of the whole SNP dataset, omitting 20% of SNPs located on A and B subgenomes only, or omitting 50% of SNPs located on D subgenome. The three analyses showed similar results thus only results of the first analysis is presented in the present paper. The idea is that if we do not have enough SNP density to cover all LD blocks in both A and B subgenomes, omitting a considerable proportion of the SNPs will mask the variance captured by the deleted SNPs while keeping the D subgenome unaffected. Obtaining the same results from the original and the masked analyses suggests that each LD block is covered with adequate number of SNPs and as such, the majority of its variance can be captured with the available SNPs.

Analysis of covariance (ANCOVA) was used to determine significant differences among the three subgenomes considering (1) the subgenome size as a covariate or (2) the chromosome size as a covariate. The fitted model for the first ANCOVA analysis was: Additive Effect ∼ subgenome + subgenome size. For the second analysis, we fitted the model twice, with and without including the interaction between chromosome size and subgenome. Thus, the models were: Additive Effect ∼ subgenome + chromosome size; and Additive Effect ∼ subgenome * chromosome size.

For each trait, a Chi-square test was performed to test whether the actual additive variance explained by the three subgenomes lies within the expected range for their values. The genome size for A, B and D subgenomes is 5727, 6274, and 4945 Mb, respectively. Thus, the expected contribution for each subgenome to the additive variance was calculated as the proportion of the subgenome size to the whole genome size, which was 33.8, 37, and 29.2% for A, B, and D subgenomes, respectively.

To further confirm that the differences among subgenomes are true and have not been inflated because of relatedness among individuals, we ran 100 replicates of the GREML analysis using randomly sampled phenotypes from the normal distribution N (0, 1). This analysis allows us to compare our findings to the null hypothesis given our data. True differences among subgenomes/chromosomal groups should be detected when using our empirical phenotypes and not simulated ones.

Finally, the reliability of the GREML analysis was estimated by running a 100 replicates of the analysis in which we omitted one random individual for each replicate (reduced model). Pearson correlation coefficients between additive variances of both models (full and reduced) for all chromosomes across all traits were computed. The reliability was estimated as the square of the average Pearson correlation coefficient over the 100 replicates. The reliability was used to calculate the "*attenuated correlation*" for all our correlation analyses following Charles (2005) implemented in Fisher (2014). Calculating the attenuated correlation avoids overestimating the significance of the correlation analysis by adjusting its value according to the standard deviation of our additive variance estimation.

RESULTS

The SHW dataset included 6,176 GBS based SNPs with missing data <20% and minor allele frequency >5%. The total heritability values ranged from 44.8 to 60.5% for resistance to Sr and SNG, respectively, (**Table 1**) with an average value of 50.4%. All estimated heritabilities were significantly higher than the heritability obtained under the null model with simulated phenotypes, which had an average of 22 and 95% confidence interval between 16.3 and 27.7%. However, it is worth noting that these values should be less than the actual heritabilities as they depend on the genotyped SNPs only (Manolio et al., 2009). The numbers presented in **Table 1** represent the proportion of the total additive variance explained by each chromosome, which sum to 100 for each trait, in which negative values were recorded as zeroes (Plotted in **Figure 1**). The original estimations and their standard deviations can be found in **Table S2**. The average standard deviation across chromosomes and traits was equal

TABLE 1 | The additive variance for different traits and its partitioning (as percentage of the total heritability) into different chromosomes, chromosomal groups, and genomes.

Chr	Size (Mb)	Al	Br	CCN	Cr	Lr	Pn	Pt	SNG	SNL	Sr	YLS	Yr
He	16,946	50.3	51.1	46.2	49.8	49.3	49.1	48.2	60.5	49.7	44.8	54.3	51.0
1A	798	3.6	7.2	3.5	10.4	6.4	0	6.8	5.1	0	0	3.9	5.7
2A	899	0	7.3	3.8	0	7.4	0	0	0.2	5.5	0	2.4	0
3A	828	7.5	4.3	6.4	0	7.2	9.0	4.7	4.7	0.8	0	1.6	9.4
4A	856	5.0	7.0	9.8	12.6	7.4	7.6	6.0	7.7	0	1.3	7.5	0
5A	827	6.1	6.3	0	0	2.6	0.8	2.1	11.4	5.2	0	0.5	3.4
6A	705	8.6	0	0	0	4.9	0	1.0	5.5	7.9	5.8	0.5	0
7A	814	6.4	0	0	6.1	5.1	2.7	3.1	5.7	0	12.4	0	0
1B	849	0	0	0	0	0	0	0	0.7	14.5	3.2	4.2	1.1
2B	928	2.8	11.4	0	3.8	2.6	6.0	7.0	0.6	7.8	18.1	7.8	1.1
3B	993	7.2	8.3	11.7	6.4	7.3	6.8	13.8	10.9	5.8	14.0	8.3	4.0
4B	821	1.4	0.4	0	11.2	3.3	7.9	8.9	1.1	9.3	15.5	7.6	2.2
5B	870	0.9	0.7	6.8	1.0	5.3	11.1	0	3.4	7.0	6.8	12.1	15.7
6B	913	9.1	10.5	7.6	3.3	4.4	10.1	0	9.2	0	0	0	0
7B	900	0	3.0	8.5	0	4.6	7.1	0	0	0	9.3	0	7.7
1D	605	10.1	12.6	6.6	8.3	2.8	3.3	5.5	10.5	8.0	0	3.8	0
2D	729	17.6	3.1	19.1	5.3	7.2	8.1	0	0	2.2	3.3	10.1	4.7
3D	771	6.9	1.9	2.1	23.5	0	3.6	6.9	9.4	7.4	1.4	9.8	7.7
4D	649	0	0	7.6	7.8	0	0	17.0	4.9	0	0	0	3.2
5D	750	0	4.1	6.6	0.1	2.3	6.3	12.9	6.9	0	4.5	0	10.2
6D	713	0.5	3.7	0	0	8.1	0	4.2	2.1	6.7	0	10.5	0
7D	728	6.1	8.4	0	0	10.9	9.6	0	0	11.7	4.6	9.4	23.8
Group1	2,252	13.8	19.8	10.1	18.7	9.2	3.3	12.3	16.3	22.5	3.2	12.0	6.8
Group2	2,556	20.5	21.7	22.9	9.2	17.2	14.1	7.0	0.7	15.5	21.4	20.4	5.9
Group3	2,592	21.5	14.5	20.2	29.9	14.5	19.4	25.4	25.0	14.1	15.5	19.7	21.1
Group4	2,326	6.5	7.4	17.3	31.6	10.7	15.5	31.9	13.7	9.3	16.7	15.1	5.3
Group5	2,447	7.0	11.1	13.4	1.1	10.2	18.2	15.0	21.7	12.3	11.3	12.6	29.3
Group6	2,331	18.2	14.2	7.6	3.3	17.4	10.1	5.2	16.8	14.6	5.8	11.0	0
Group7	2,442	12.5	11.4	8.5	6.1	20.6	19.4	3.1	5.7	11.7	26.2	9.4	31.5
A	5,727	37.2	32.2	23.5	29.1	41.0	20.2	23.8	40.3	19.5	19.4	16.5	18.6
B	6,274	21.5	34.1	34.6	25.7	27.6	48.9	29.8	25.8	44.5	66.8	39.9	31.8
D	4,945	41.4	33.7	41.9	45.2	31.4	31.0	46.4	33.9	36.0	13.8	43.6	49.6
Chi test	–	0.003	NS	0.01	0.002	NS	0.009	0.001	NS	0.01	0	0	0

*Negative estimations were set to 0 in this table but detailed information can be found in **Table S2**. The last row represents Chi square p-value which compares the actual fractional contribution of A, B, and D subgenomes to the additive variance with the expected one which assumes the percentage of the subgenome size, 33.8, 37, and 29.2% for A, B, and D subgenomes, respectively. NS: not significant at 0.05.*

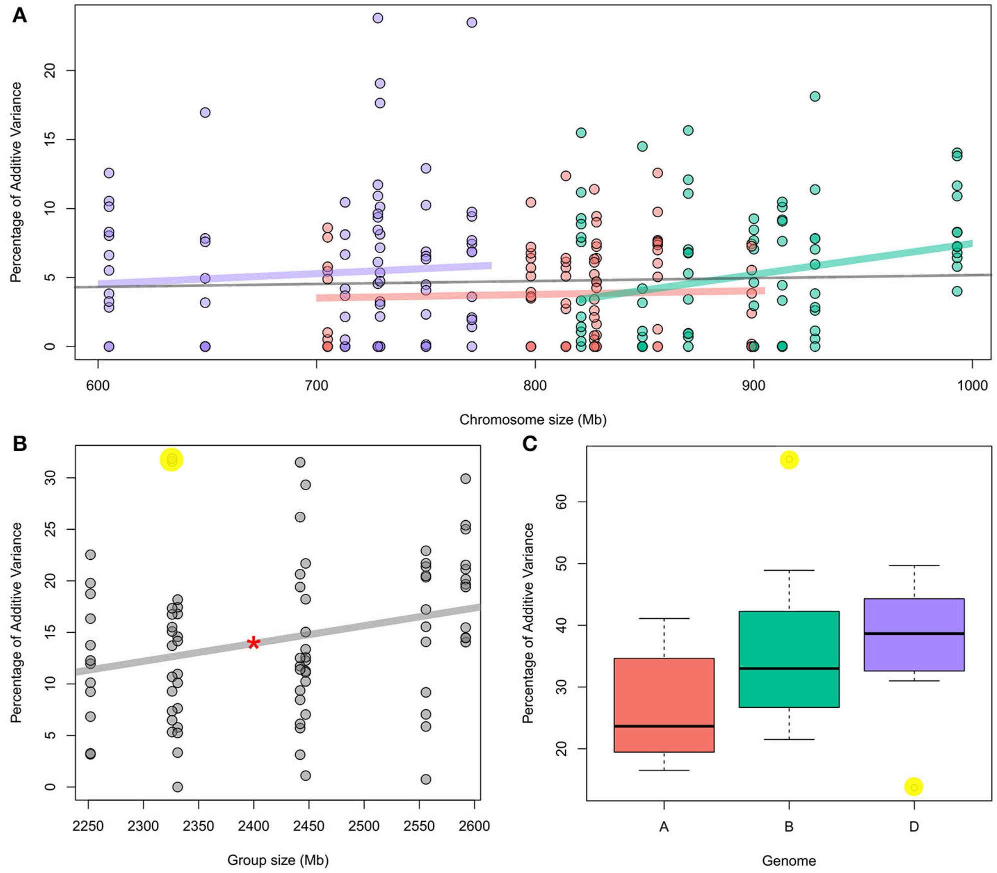


FIGURE 1 | (A) Percentage of individual chromosome contribution to the additive variance of 12 traits as function to chromosome size; red: "A" genome chromosomes; Green: "B" genome chromosomes; and Purple: "D" genome chromosomes. The gray line represents the correlation for all 21 chromosomes. For individual traits, see **Figure S1**. **(B)** Percentage of each chromosomal group (seven groups) contribution to the additive variance of 12 traits as function to chromosome size. Red star over the correlation line represents its significance at $P < 0.05$. For individual traits, see **Figure S2**. **(C)** Boxplot showing the contribution of each genome to the additive variance of 12 traits. Highlighted yellow dots in b and c represent the outliers. For detail information, see **Table 1**.

to 0.077 while the reliability of the GREML analysis given the standard deviation was equal to 0.45 (0.67^2). The considerably low reliability is a result of small population size and relatedness among individuals.

For the 21 chromosomes across all traits, we found no correlation between chromosome sizes and their explained additive variance (**Figure 1A**; **Table 2**). However, for individual traits, only Sr resistance showed a significant correlation between all 21 chromosomes and their fractional contribution to the additive variance with p-value = 0.04 and r = 0.45 (**Table 2**; **Figure S1**). The median r value between chromosome size and fractional additive variance for all traits was equal to 0.005. When chromosomes within each subgenome were considered, only the additive variance explained by the B subgenome chromosomes showed a significant but weak correlation with chromosome size (p-value = 0.02 and r = 0.25; **Figure 1A**; **Table 2**). Neither the Sr correlation nor the B subgenome correlation were significant after adjusting them for attenuation following Charles (2005).

A significant correlation was evident between the cumulative size for each chromosomal group and the fractional additive variance explained by the group with p-value = 0.01 and r = 0.27 (**Figure 1B**, **Table 2**). Removing two outliers (the contribution of group 4 for Cr and Pt resistance which are highlighted in yellow, **Figure 1B**) strengthened this correlation with p-value = 0.001 and r = 0.34. However, when correcting the correlation for attenuation, it was significant only after removing the two outliers with p-value = 0.037 and r = 0.23. A single chromosome with the highest contribution within each group can explain about 69% of the total group additive variance on average across all traits. The relationship between fractional additive variance and the chromosomal group cumulative size for individual traits had a median value of 0.43 (**Table 2**) and is plotted in **Figure S2**.

The cumulative fractional additive variance significantly varied between the three subgenomes. The median values for the percentage of additive variance contributed by A, B, and D subgenomes were 23.7, 33, and 38.7%; respectively (**Figure 1C**). These values changed to 23.8, 31.8, and 41.3%, respectively, after omitting stem rust resistance, an outlier compared to other traits. ANCOVA analysis that considered the genome size as a covariate confirmed the significant differences among the three

TABLE 2 | Pearson correlation coefficient (*r* values) between the additive variance explained by all 21 chromosome sizes (column All), chromosomes within each subgenome (A, B, and D) and chromosomal group size (Groups).

Stress	A	B	D	Groups	All
SNL	−0.329	−0.435	0.007	−0.273	−0.067
SNG	−0.240	0.668	−0.283	−0.084	−0.045
YLS	0.382	−0.024	0.438	0.690	0.054
Cr	0.121	−0.049	0.119	0.002	−0.118
Lr	0.426	0.628	0.176	0.438	0.139
Sr	−0.389	0.208	0.620	0.597	$0.451^{*\dagger\dagger}$
Yr	0.006	−0.093	0.474	0.413	−0.062
CCN	0.502	0.671	−0.110	0.674	0.067
Pn	0.270	0.126	0.428	0.700	0.336
Pt	0.032	0.463	−0.223	0.004	−0.134
Br	0.683	0.764	−0.463	0.189	0.182
Al	−0.742	0.654	−0.022	0.536	−0.204
Combined	0.0407	0.25*	0.075	$0.27^{*\dagger\dagger}$ $(0.34^{**\dagger})$	0.043

The final row represents the r values considering all traits together (visualized in ***Figures 1A,B****).* †*Represents the correlation coefficient after removing the two outliers in* ***Figure 1B****; this was significant at p-value < 0.05 after correcting for attenuation. *Significant at p-value < 0.05; **Significant at p-value < 0.01.* ††*Not significant after correcting for attenuation.*

subgenomes across all 12 traits with p-values = 0.01. This was the only significant component in the model. The ANCOVA analysis that considered the size of chromosomes as a covariate had a p-value of 0.006 (same value with and without including the interaction between genome and chromosome size in the model) which was the only significant component in both models.

For individual traits, Chi-square tests showed significant differences between the actual and the expected subgenome contribution to all traits except for Br, Lr, and SNG. For Al, CCN, Cr, Pt, and Yr, only the contribution of the D subgenome was higher than expected, while the contributions of the B and D subgenomes were higher than expected for Pn, SNL, and YLS (**Table 1**). Br, Lr, and SNG resistances were not significantly different from the expected contribution, but the actual contribution of the D subgenome for all of them was slightly higher than expected (**Table 1**).

Population structure, linkage disequilibrium, and relatedness among individuals did not have an effect on our results. The inclusion of the first 10 principal components as covariates in the model did not have a large effect on heritability estimates (data not shown) which means that population structure has minimal effect on the heritability estimations. Similarly, further analysis with a randomly chosen subset of SNPs did not affect the results either (**Table S3**), indicating that the extended linkage disequilibrium observed in the D subgenome in this population did not overestimate the contribution of the D subgenome. Furthermore, under the null hypothesis using simulated phenotypes, the cumulative additive variance was 0.0698 (±0.026), 0.0735 (±0.027) and 0.0766 (±0.029) for the A, B, and D subgenomes, respectively, indicating true differences among subgenomes observed with empirical phenotypes that are not affected by relatedness among individuals.

Estimating the conditional effect for each chromosome based on the other 20 chromosomes showed considerable correlation among chromosomes (**Table 3**; **Table S2**). On average for all chromosomes across all traits, 46% of chromosome additive variance can be explained by other chromosomes. This value ranged from 20.6% for Yr resistance to 57.3% for Br tolerance (Inferred from **Table 3**). Interestingly, even for the conditional analysis after excluding correlated additive variances, our conclusion that the D genome had the highest contribution to the total heritability did not change with 22.3, 31.9, and 44.8% of the total additive variance attributed to the A, B, and D subgenomes, respectively. Removing Sr increased the D subgenome contribution to 45.7% and reduced the B subgenome contribution to 30.1%. The correlation among all 21 GRMs also support these results (**Figure 2**). All GRMs for the A and B subgenome chromosomes clustered together while GRMs for D subgenome chromosomes formed another cluster. Thus, the correlated additive variance can be explained by the same ancestor supporting the superiority of the D subgenome regardless of the low reliability of the GREML analysis.

DISCUSSION

Decomposing additive genetic variance based on different set of SNPs has become a commonly used method in quantitative genetics in recent years (Yang et al., 2010, 2011a,b, 2015; Lee et al., 2012). Researchers usually remove related individuals to ensure that they are capturing SNP-based heritability only (Yang et al., 2017). Although this is possible in human genetics and some animal populations that have large effective population size, it is impossible to have such optimal populations containing distinctly related individuals in species such as bread wheat with extremely small effective population sizes (Joukhadar et al., 2017). For this reason, the heritability estimated with this method in populations of species such as bread wheat will be a mixture of SNP-based heritability from phenotypic correlation due to unrelated individuals and pedigree-based heritability from phenotypic correlation due to relatedness (Yang et al., 2017). One advantage of using related individuals is that the analysis requires smaller populations to obtain an acceptable standard error (SE), because SE is negatively correlated with the average relatedness among individuals. Yang et al. (2017) pointed out that the SE can be further decreased if rare SNPs are excluded from the analysis.

Linkage disequilibrium (LD) can cause a huge bias for decomposing additive variance analysis as the variance estimation depends on the LD between the causal variant and the closest genotyped SNPs (Speed et al., 2012). The D subgenome in our population showed large LD blocks (Jighly et al., 2016) but this did not result in over estimating its contribution because there were sufficient SNPs to capture most additive variance in the A and B subgenomes (**Table S3**). This is not unexpected for populations with small effective population size like SHW. For example, randomly selecting 10K out of 354K SNPs reduced the captured additive variance by only 1% for different traits in chickens (Abdollahi-Arpanahi et al., 2014). Population structure also did not affect the estimation as the

TABLE 3 | The heritability estimation using the conditional effect model (excluding the GRM of one chromosome).

Trait	Al	Br	CCN	Cr	Lr	Pn	Pt	SNG	SNL	Sr	YLS	Yr
He	0.503	0.511	0.462	0.498	0.493	0.491	0.482	0.605	0.497	0.448	0.543	0.510
1A	0.49 (0.013)	0.507 (0.004)	0.455 (0.007)	0.472 (0.027)	0.48 (0.013)	0.49 (0.001)	0.469 (0.013)	0.592 (0.013)	0.508 (−0.011)	0.462 (−0.014)	0.544 (−0.001)	0.486 (0.024)
2A	0.491 (0.012)	0.504 (0.007)	0.453 (0.009)	0.489 (0.009)	0.475 (0.018)	0.485 (0.007)	0.484 (−0.002)	0.607 (−0.002)	0.49 (0.007)	0.461 (−0.013)	0.541 (0.002)	0.515 (−0.005)
3A	0.486 (0.018)	0.499 (0.012)	0.451 (0.011)	0.491 (0.007)	0.481 (0.012)	0.474 (0.017)	0.477 (0.005)	0.597 (0.008)	0.499 (−0.002)	0.446 (0.002)	0.543 (0.001)	0.475 (0.035)
4A	0.487 (0.016)	0.506 (0.005)	0.432 (0.03)	0.465 (0.034)	0.484 (0.01)	0.478 (0.013)	0.476 (0.006)	0.589 (0.016)	0.492 (0.005)	0.448 (0)	0.528 (0.016)	0.512 (−0.002)
5A	0.494 (0.009)	0.496 (0.015)	0.464 (−0.002)	0.504 (−0.006)	0.496 (−0.003)	0.49 (0.001)	0.479 (0.003)	0.586 (0.02)	0.49 (0.007)	0.455 (−0.007)	0.545 (−0.001)	0.502 (0.008)
6A	0.483 (0.02)	0.499 (0.012)	0.458 (0.004)	0.49 (0.008)	0.484 (0.009)	0.487 (0.004)	0.487 (−0.005)	0.593 (0.012)	0.485 (0.012)	0.433 (0.015)	0.552 (−0.009)	0.51 (0)
7A	0.487 (0.016)	0.474 (0.037)	0.451 (0.011)	0.49 (0.008)	0.485 (0.008)	0.482 (0.009)	0.476 (0.006)	0.599 (0.006)	0.496 (0.001)	0.426 (0.022)	0.537 (0.006)	0.505 (0.005)
1B	0.519 (−0.016)	0.502 (0.009)	0.471 (−0.009)	0.508 (−0.01)	0.489 (0.005)	0.481 (0.01)	0.494 (−0.012)	0.612 (−0.007)	0.471 (0.026)	0.451 (−0.003)	0.544 (−0.001)	0.505 (0.005)
2B	0.496 (0.007)	0.463 (0.048)	0.445 (0.017)	0.492 (0.006)	0.491 (0.002)	0.474 (0.017)	0.478 (0.004)	0.607 (−0.002)	0.48 (0.017)	0.408 (0.04)	0.517 (0.026)	0.508 (0.002)
3B	0.483 (0.02)	0.51 (0.001)	0.431 (0.031)	0.47 (0.028)	0.474 (0.019)	0.483 (0.008)	0.458 (0.024)	0.585 (0.02)	0.48 (0.017)	0.423 (0.025)	0.526 (0.018)	0.509 (0.001)
4B	0.51 (−0.007)	0.514 (−0.003)	0.473 (−0.011)	0.437 (0.061)	0.482 (0.011)	0.472 (0.019)	0.471 (0.011)	0.605 (0)	0.473 (0.025)	0.418 (0.03)	0.523 (0.02)	0.472 (0.038)
5B	0.502 (0.001)	0.513 (−0.002)	0.455 (0.007)	0.497 (0.001)	0.48 (0.013)	0.473 (0.018)	0.487 (−0.005)	0.591 (0.014)	0.496 (0.001)	0.44 (0.008)	0.506 (0.037)	0.469 (0.042)
6B	0.471 (0.033)	0.508 (0.003)	0.443 (0.02)	0.486 (0.012)	0.487 (0.006)	0.462 (0.029)	0.484 (−0.002)	0.585 (0.02)	0.496 (0.001)	0.456 (−0.008)	0.514 (0.029)	0.5 (0.01)
7B	0.5 (0.003)	0.514 (−0.003)	0.445 (0.017)	0.493 (0.005)	0.488 (0.005)	0.487 (0.004)	0.482 (0)	0.626 (−0.021)	0.498 (−0.001)	0.431 (0.017)	0.555 (−0.012)	0.496 (0.015)
1D	0.47 (0.033)	0.52 (−0.009)	0.454 (0.008)	0.47 (0.028)	0.496 (−0.003)	0.483 (0.008)	0.475 (0.007)	0.547 (0.058)	0.48 (0.017)	0.438 (0.01)	0.532 (0.011)	0.484 (0.027)
2D	0.463 (0.04)	0.511 (0)	0.364 (0.098)	0.483 (0.015)	0.472 (0.021)	0.475 (0.017)	0.465 (0.017)	0.581 (0.024)	0.494 (0.003)	0.444 (0.004)	0.528 (0.015)	0.486 (0.024)
3D	0.49 (0.013)	0.506 (0.005)	0.467 (−0.005)	0.401 (0.097)	0.48 (0.013)	0.484 (0.007)	0.471 (0.011)	0.592 (0.013)	0.474 (0.023)	0.442 (0.006)	0.511 (0.032)	0.486 (0.024)
4D	0.504 (−0.001)	0.494 (0.017)	0.445 (0.017)	0.481 (0.017)	0.489 (0.004)	0.491 (0)	0.429 (0.053)	0.589 (0.016)	0.491 (0.006)	0.456 (−0.008)	0.534 (0.009)	0.498 (0.012)
5D	0.49 (0.013)	0.494 (0.017)	0.454 (0.008)	0.497 (0.001)	0.489 (0.004)	0.481 (0.011)	0.446 (0.036)	0.589 (0.016)	0.483 (0.014)	0.431 (0.017)	0.503 (0.04)	0.455 (0.055)
6D	0.507 (−0.004)	0.504 (0.007)	0.454 (0.008)	0.494 (0.004)	0.47 (0.023)	0.494 (−0.003)	0.473 (0.01)	0.607 (−0.002)	0.477 (0.02)	0.443 (0.005)	0.514 (0.029)	0.497 (0.013)
7D	0.491 (0.012)	0.493 (0.019)	0.457 (0.005)	0.498 (0)	0.469 (0.024)	0.47 (0.021)	0.478 (0.004)	0.589 (0.017)	0.469 (0.029)	0.443 (0.005)	0.515 (0.028)	0.445 (0.065)
A contribution%	37.3	42.2	23.4	25.3	31.8	23.5	15.7	27.5	13.9	18.9	7.8	17.8
B contribution%	22.9	28.0	29.9	30.7	27.7	47.5	18.6	19.8	37.7	58.3	40.8	27.9
D contribution%	39.8	29.8	46.8	44.0	40.5	29.0	65.7	52.7	48.5	22.8	51.4	54.3
Chi test	0.008	NS	0.001	0.005	0.03	0.05	0	0	0	0	0	0

The values between brackets describes the additive variance inferred from the full model (the first row in the table) minus the conditional total additive variance. The last three rows represent the contribution of each subgenome to total independent additive variance (values between brackets). The last row represents Chi square p-value which compares the conditional contribution of A, B and D subgenomes to the additive variance with the expected one which assumes the percentage of the subgenome size, 33.8, 37, and 29.2% for A, B, and D subgenomes, respectively. NS, not significant at 0.05.

estimations were very similar to the model that involved the first 10 PCs as covariates (Lee et al., 2012), although considerable correlation between different chromosomes was observed in this germplasm (**Table 3**; **Table S2**). On the other hand, this correlation did not affect our conclusion that the D subgenome had a higher contribution to the total additive variance relative

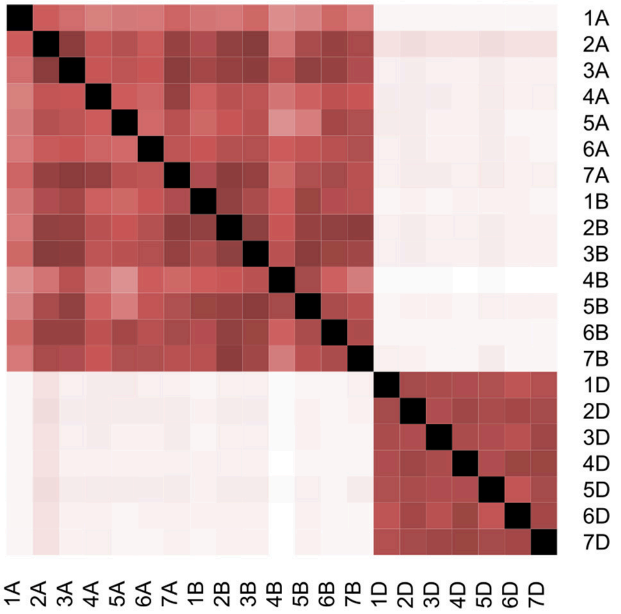


FIGURE 2 | Pairwise correlation between all 21 GRMs for wheat chromosomes. White color represents Pearson coefficient = 0, while black color represents Pearson coefficient = 1.

to the A and B subgenomes (**Table 3**; **Table S2**), and especially that GRMs of the D subgenome chromosomes were clustered together and were not correlated with any of the 14 GRMs of the A and B subgenome chromosomes (**Figure 2**).

Almost all studies that have partitioned additive variance have shown a significant correlation exists between chromosome size and variance (e.g., Yang et al., 2011b; Lee et al., 2012; Robinson et al., 2013). In the present study using SHW, however, chromosome size was not correlated with explained additive variance for any trait, although a weak correlation was observed for chromosomes within the B subgenome. The significant correlation for Sr (**Table 2**) cannot be attributed to chromosome size directly, but rather to differences in size between D and B subgenomes, which explained 13.8 and 66.8% of the additive variance, respectively (**Figure S1**; **Table 1**). The previous two correlations became non-significant after correcting for attenuation.

In contrast to what we found for all individual chromosomes, a significant but weak correlation was found between the cumulative sizes and cumulative additive variances for each chromosomal group (**Figure 1B**). In polyploids, the balanced dosage hypothesis, which involves gene loss, functional divergence and epigenetic changes in newly synthesized polyploids, has been widely discussed and has been proven for many gene families (Ohno, 1970; Lynch and Conery, 2000; Tate et al., 2009; Buggs et al., 2010, 2012; Xiong et al., 2011; Feldman and Levy, 2012; Conant et al., 2014; Dodsworth et al., 2016). We hypothesize that these structural and functional changes during diploidization keep a single functional copy for each gene in one homoeolog and thus, larger chromosomes may not necessarily have higher contribution to the additive variance if functional copies are not distributed equally in the three homoeologs. Instead, when considering the three homoeologs together, all genes will have functional copies. Thus, larger chromosomal groups may have higher contribution to the additive variance. This may explain the correlation between group size and effect. Another important finding is that one homoeolog can dominate the group additive effect within each chromosomal group with an average of 69% of the total group additive variance (Inferred from **Table 1**). Future research using larger populations should consider the relation between variance and chromosome size in both SHWs and their progenitors to further confirm this finding and to better understand underlying mechanisms that allow one homoeolog to dominate the group additive effect.

Pont et al. (2013) showed that the D subgenome generally dominated the tetraploid A and B subgenomes in hexaploid wheat by analyzing synteny and conserved orthologous gene data. Our results also showed this for stress resistance traits and that the dominance effect of the D subgenome was greater with regard to the A than the B subgenome with the median percentage of additive variance across all traits for A subgenome being 23.7% (**Figure 1C**). However, this cannot be generalized for all traits. For instance, the A subgenome contributed 9.6% more than the D subgenome to Lr resistance, whereas the B subgenome dominated the A and D subgenomes for Sr resistance (**Table 1**). Lagudah et al. (1993) showed that transferring Sr and Lr resistance form *Ae. tauschii* to hexaploid wheat is partially or fully suppressed by unknown mechanisms while Kerber and Green (1980) reported a suppressor for A and B subgenome Sr resistance in chromosome 7D. Later studies have indicated that suppression of the resistance of one subgenome of bread wheat by the other subgenomes is affected by SHW parents and pathogen isolates (Kema et al., 1995; Badebo et al., 1997; Ogbonnaya et al., 2013). Thus, efficient implementation of SHW in breeding programs should combine superior chromosomes within each chromosomal group for each trait independently, although the general trend showed that the D subgenome had a higher contribution to the additive variance. Future research should investigate suppression mechanisms and whether the general D subgenome superior additive contribution is a result of suppressing A and B subgenomes resistance to different biotic and abiotic stresses.

AUTHOR CONTRIBUTIONS

AJ: suggested and planned the study, analyzed the data and drafted the manuscript; RJ: assisted with R scripting and drafted the manuscript; SS: provided the GBS data; FO: planned the study, provided the phenotypic data, drafted the manuscript and gave the final acceptance for the manuscript to be submitted; All authors read and approved the final copy of the manuscript.

ACKNOWLEDGMENTS

We thank Seeds of Discovery—Sustainable Modernization of Traditional Agriculture project (MasAgro), Mexico for supporting the genotyping work. The Grains Research and Development Corporation funded Synthetic Evaluation Project in Australia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00027/full#supplementary-material

Figure S1 | Percentage of individual chromosome contribution to the additive variance for each trait as function to chromosome size. Colors represents different subgenomes; red: "A" subgenome chromosomes; Green: "B" subgenome chromosomes; and Purple: "D" subgenome chromosomes. The gray line represents the correlation for all 21 chromosomes.

Figure S2 | Percentage of chromosomal group contribution to the additive variance for each trait as function to chromosome size.

Table S1 | Pedigree and passport information for the SHW population.

Table S2 | The first line for each chromosome contains information about the estimated additive variance for different traits and their standard deviations, between brackets, using the full model (the model that fits 21 GRMs). The second line for each chromosome is the heritability estimation using the conditional effect model (excluding the GRM of one chromosome). Values between brackets describes the additive variance inferred from the full model (the first row in the table) minus the conditional total additive variance. The second line is exactly the same as **Table 3** in the paper but was repeated here for easier comparisons between the full and the conditional models.

Table S3 | The additive variance for different traits and its partitioning (as percentage of the total heritability) into different chromosomes, chromosomal groups and subgenomes for subset of the whole data set that includes 80% of our SNPs.

REFERENCES

Abdollahi-Arpanahi, R., Pakdel, A., Nejati-Javaremi, A., Moradi Shahrbabak, M., Morota, G., Valente, B. D., et al. (2014). Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. *J. Anim. Breed. Genet.* 131, 183–193. doi: 10.1111/jbg.12079

Badebo, A., Kema, G. H. J., van Ginkel, M., and van Silfhout, C. H. (1997). Genetics of suppressors of resistance to stripe rust in synthetic wheat hexaploids derived from *Triticum turgidum* subsp. *dicoccoides* and *Aegilops squarrosa*. *Afr. Crop Sci. Conf. Proc.* 3, 195–202.

Buggs, R. J., Chamala, S., Wu, W., Tate, J. A., Schnable, P. S., Soltis, D. E., et al. (2012). Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.* 22, 248–252. doi: 10.1016/j.cub.2011.12.027

Buggs, R. J., Elliott, N. M., Zhang, L., Koh, J., Viccini, L. F., Soltis, D. E., et al. (2010). Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytol.* 186, 175–183. doi: 10.1111/j.1469-8137.2010.03205.x

Charles, E. P. (2005). The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets. *Psychol. Methods* 10, 206–226. doi: 10.1037/1082-989X.10.2.206

Chen, G. B., Lee, S. H., Brion, M. J., Montgomery, G. W., Wray, N. R., Radford-Smith, G. L., et al. (2014). Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* 23, 4710–4720. doi: 10.1093/hmg/ddu174

Conant, G. C., Birchler, J. A., and Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* 19, 91–98. doi: 10.1016/j.pbi.2014.05.008

Dodsworth, S., Chase, M. W., and Leitch, A. R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Bot. J. Linn. Soc.* 180, 1–5. doi: 10.1111/boj.12357

Dreccer, F. M., Borgognone, G. M., Ogbonnaya, F. C., Trethowan, R. M., and Winter, B. (2007). CIMMYT-selected derived synthetic bread wheats for rainfed environments: yield evaluation in Mexico and Australia. *Field Crops Res.* 100, 218–228. doi: 10.1016/j.fcr.2006.07.005

Dreccer, F. M., Ogbonnaya, F. C., and Borgognone, M. G. (2004). "Sodium exclusion in primary synthetic wheats," in *Proc. XI Wheat Breeding Assembly* (Canberra, CT), 118–121.

El-Bouhssini, M., Ogbonnaya, F. C., Chen, M., Lhaloui, S., Rihawi, F., and Dabbous, A. (2013). Sources of resistance in primary synthetic hexaploid wheat (*Triticum aestivum* L.) to insect pests: Hessian fly, Russian wheat aphid and Sunn pest in the fertile crescent. *Genet. Resour. Crop Evol.* 60, 621–627. doi: 10.1007/s10722-012-9861-3

Emebiri, L. C., and Ogbonnaya, F. C. (2015). Exploring the synthetic hexaploid wheat for novel sources of tolerance to excess boron. *Mol. Breed.* 35:68. doi: 10.1007/s11032-015-0273-x

Feldman, M., and Levy, A. A. (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* 192, 763–774. doi: 10.1534/genetics.112.146316

Fisher, C. R. (2014). A pedagogic demonstration of attenuation of correlation due to measurement error. *Spreadsheets Educ.* 7:4. Available online at: http://epublications.bond.edu.au/ejsie/vol7/iss1/4/

Goldstein, D. B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.* 360, 1696–1698. doi: 10.1056/NEJMp0806284

Gororo, N. N., Eagles, H. A., Eastwood, R. F., Nicolas, M. E., and Flood, R. G. (2002). Use of *Triticum tauschii* to improve yield of wheat in low-yielding environments. *Euphytica* 123, 241–254. doi: 10.1023/A:1014910000128

Hegarty, M. J., Barker, G. L., Brennan, A. C., Edwards, K. J., Abbott, R. J., and Hiscock, S. J. (2008). Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in Senecio. *Phil. Trans. R. Soc. B* 363, 3055–3069. doi: 10.1098/rstb.2008.0080

Jighly, A., Alagu, M., Makdis, F., Singh, M., Singh, S., Emebiri, L. C., et al. (2016). Genomic regions conferring resistance to multiple fungal pathogens in synthetic hexaploid wheat. *Mol. Breed.* 36:127. doi: 10.1007/s11032-016-0541-4

Joukhadar, R., Daetwyler, H. D., Bansal, U. K., Gendall, A. R., and Hayden, M. J. (2017). Genetic diversity, population structure and ancestral origin of Australian wheat. *Front. Plant Sci.* 8:2115. doi: 10.3389/fpls.2017.02115

Joukhadar, R., El-Bouhssini, M., Jighly, A., and Ogbonnaya, F. C. (2013). Genomic regions associated with resistance to five major pests in wheat. *Mol. Breed.* 32, 943–960. doi: 10.1007/s11032-013-9924-y

Kashkush, K., Feldman, M., and Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160, 1651–1659. Available online at: http://www.genetics.org/content/160/4/1651.long

Kema, G. H. J., Lange, L., and van Silfhout, C. H. (1995). Differential suppression of stripe rust resistance in synthetic wheat hexaploids derived from *Triticum turgidum* subsp. *dicoccoides* and *Aegilops squarrosa*. *Phytopathology* 85, 425–429.

Kerber, E. R., and Green, G. J. (1980). Suppression of stem rust resistance in the hexaploid wheat cv. Canthatch by chromosome 7DL. *Can. J. Bot.* 58, 1347–1350.

Lagudah, E. S., Appels, R., McNeil, D., and Schachtman, D. P. (1993). "Exploiting the diploid D genome chromatin for wheat improvement," in *Gene Conservation and Exploitation,* eds J. P. Gustafson, R. Appels, and P. Raven (New York, NY: Plenum Press), 87–107.

Lee, S. H., and van der Werf, J. H. (2016). MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* 32, 1420–1422. doi: 10.1093/bioinformatics/btw012

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), International Schizophrenia Consortium (ISC), et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* 44, 247–250. doi: 10.1038/ng.1108

Lee, S. H., Harold, D., Nyholt, D. R., ANZGene Consortium, International Endogene Consortium, Genetic and Environmental Risk for Alzheimer's disease Consortium, et al. (2013). Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum. Mol. Genet.* 22, 832–841. doi: 10.1093/hmg/dds491

Lynch, M., and Conery, J. S. (2000). The evolutionary fate of duplicated genes. *Science* 290, 1151–1154. doi: 10.1126/science.290.5494.1151

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Mulki, M. A., Jighly, A., Ye, G. Y., Emebiri, L. C., Moody, D., Ansari, O., et al. (2013). Association mapping for soilborne pathogen resistance in synthetic hexaploid wheat. *Mol. Breed.* 31, 299–311. doi: 10.1007/s11032-012-9790-z

Ogbonnaya, F. C., Abdalla, O., Mujeeb-Kazi, A., Kazi, A. G., Xu, S. S., Gosman, N., et al. (2013). Synthetic hexaploids: harnessing species of the primary gene pool for wheat improvement. *Plant Breed. Rev.* 37, 35–122. doi: 10.1002/9781118497869.ch2

Ogbonnaya, F. C., Huang, S., Steadman, E., Emebiri, L. C., Dreccer, M. F., Lagudah, E., et al. (2008a). "Mapping quantitative trait loci associated with salinity tolerance in synthetic derived backcrossed bread lines," in *Proceedings of 11th International Wheat Genetics Symposium*, eds R. Appels, R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. McIntyre, and P. Sharp (Brisbane, QLD: Sydney University Press).

Ogbonnaya, F. C., Imtiaz, M., Bariana, H. S., McLean, M., Shankar, M., Hollaway, G. J., et al. (2008b). Mining synthetic hexaploids for multiple disease resistance to improve wheat. *Aust. J. Agric. Res.* 59, 421–431. doi: 10.1071/AR07227

Ogbonnaya, F. C., Ye, G., Trethowan, R., Dreccer, F., Lush, D., Shepperd, J., et al. (2007). Yield of synthetic backcross-derived lines in rainfed environments of Australia. *Euphytica* 157, 321–336. doi: 10.1007/s10681-007-9381-y

Ohno, S. (1970). *Evolution by Gene Duplication.* New York, NY: Springer-Verlag.

Ozkan, H., Levy, A. A., and Feldman, M. (2001). Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops–Triticum*) group. *Plant Cell* 13, 1735–1747. doi: 10.1105/tpc.13.8.1735

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

Raman, H., Stodart, B., Ryan, P. R., Delhaize, E., Emebiri, L., Raman, R., et al. (2010). Genome wide association analyses of common wheat (*Triticum aestivum* L) germplasm identifies multiple loci for aluminum resistance. *Genome* 53, 957–966. doi: 10.1139/G10-058

Robinson, M. R., Santure, A. W., DeCauwer, I., Sheldon, B. C., and Slate, J. (2013). Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. *Mol. Ecol.* 22, 3963–3980. doi: 10.1111/mec.12375

Sehgal, D., Vikram, P., Sansaloni, C. P., Ortiz, C., Pierre, C. S., Payne, T., et al. (2015). Exploring and mobilizing the gene bank biodiversity for wheat improvement. *PLoS ONE* 10:e0132112. doi: 10.1371/journal.pone.0132112

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* 13, 1749–1759. doi: 10.1105/tpc.13.8.1749

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* 91, 1011–1021. doi: 10.1016/j.ajhg.2012.10.010

Stace, C. A. (1987). Triticale: a case of nomenclatural mistreatment. *Taxon* 36, 445–452. doi: 10.2307/1221447

Tate, J. A., Joshi, P., Soltis, K. A., Soltis, P. S., and Soltis, D. E. (2009). On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biol.* 9:80. doi: 10.1186/1471-2229-9-80

Weiss-Schneeweiss, H., Emadzade, K., Jang, T. S., and Schneeweiss, G. M. (2013). Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenet. Genome Res.* 140, 137–150. doi: 10.1159/000351727

Xiong, Z., Gaeta, R. T., and Pires, J. C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl. Acad. Sci. U.S.A.* 108. 7908–7913. doi: 10.1073/pnas.1014138108

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi: 10.1038/ng.3390

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525. doi: 10.1038/ng.823

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* 49, 1304–1310. doi: 10.1038/ng.3941

Yang, W., Liu, D., Li, J., Zhang, L., Wei, H., Hu, X., et al. (2009). Synthetic hexaploid wheat and its utilization for wheat genetic improvement in China. *J. Genet. Genomics* 36, 539–546. doi: 10.1016/S1673-8527(08)60145-9

Zegeye, H., Rasheed, A., Makdis, F., Badebo, A., and Ogbonnaya, F. C. (2014). Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. *PLoS ONE* 9:e105593 doi: 10.1371/journal.pone.0105593

Conflict of Interest Statement: 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.