ADVANCES IN APPLIED BIOINFORMATICS IN CROPS

EDITED BY: Mary-Ann Blätke, Sebastian Beier, Uwe Scholz, Evgeny Gladilin and Jedrzej Jakub Szymanski

PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-620-1
DOI 10.3389/978-2-88966-620-1

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ADVANCES IN APPLIED BIOINFORMATICS IN CROPS

Topic Editors:
Mary-Ann Blätke, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Sebastian Beier, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Uwe Scholz, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Evgeny Gladilin, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Jedrzej Jakub Szymanski, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Citation: Blätke, M.-A., Beier, S., Scholz, U., Gladilin, E., Szymanski, J. J., eds. (2021). Advances in Applied Bioinformatics in Crops. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-620-1

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 12 February 2021
doi: 10.3389/fpls.2021.640394

Editorial: Advances in Applied Bioinformatics in Crops

Mary-Ann Blätke[1], *Jedrzej Jakub Szymanski*[1], *Evgeny Gladilin*[1], *Uwe Scholz*[2] *and Sebastian Beier*[2]*

[1] Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany,
[2] Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany

Keywords: breeding informatics, image analysis, data visualization, biodiversity, plant (multi)omics, computational biology, high-throughput technologies in plants

Editorial on the Research Topic

Advances in Applied Bioinformatics in Crops

INTRODUCTION

Big Data in life science is scattered across hundreds of unstructured data sets, biological databases and thousands of scientific journals. Modern crop research relies on high-throughput technologies that generate large quantities of high-dimensional data. The challenge for Applied Bioinformatics is to capture, model, integrate, analyze, visualize and make these data accessible in a FAIR (Wilkinson et al., 2016) (https://fair-dom.org) way. This, in turn, translates directly to the improvement of our understanding of crop biology, and in practical terms results in the development of new elite genotypes and improvement of plant cultivation strategies.

The presented collection of articles describes the flow of information from high-throughput data acquisition, data processing and analysis, the underlying IT infrastructure, and the modeling of biological processes. This Research Topic (RT) special issue is based on contributions to the Fifteenth *Gatersleben Research Conference on Applied Bioinformatics in Crops* carried out during March 2019 at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, Germany, and includes contributions from scientists who are researching related topics. This conference was of interest to Life Scientists, Bioinformaticians, Computer Scientists, Systems Biologists, Synthetic Biologists, and others working or interested in the developing area of Applied Bioinformatics for crops.

Advances in high-throughput technologies, such as next-generation sequencing (NGS) have paved the way for turning life sciences into data-intensive disciplines. NGS has enabled leaps forward in plant genomics by tremendously increasing the number of sequenced genomes and assembling pan genomes to explore genomic diversity (Bayer et al., 2020). Other disciplines such as genome-wide association studies (GWAS), transcriptomic, proteomic, and metabolomic profiling benefit from advanced high-throughput experimental techniques and other omic-related fields. Those advances would not have been possible without bioinformatics providing novel tools, databases, and the other resources required to analyze the ever-increasing amounts of data. Nevertheless, extracting the inherent valuable knowledge hidden within numerous, large and diverse data sets remains a daunting challenge in bioinformatics and computational biology (Tang et al., 2019). In this respect, deep learning advances the bioinformatics toolkit and is of unprecedented value to unveil precious insights linked to genetic information, molecules and molecular processes. Deep learning in plant sciences allows us to conduct high-throughput phenotyping based on classical image data, as well as complex comparative genomic, transcriptomic, and proteomic studies, see Soltis et al. (2020) and Tang et al. (2019) for a review.

OPEN ACCESS

Edited by:
Wagner L. Araújo,
Universidade Federal de Viçosa, Brazil

Reviewed by:
João Henrique Frota Cavalcanti,
Federal University of Amazonas, Brazil

***Correspondence:**
Sebastian Beier
beiers@ipk-gatersleben.de

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *11 December 2020*
Accepted: *20 January 2021*
Published: *12 February 2021*

Citation:
Blätke M-A, Szymanski JJ, Gladilin E, Scholz U and Beier S (2021) Editorial: Advances in Applied Bioinformatics in Crops. Front. Plant Sci. 12:640394. doi: 10.3389/fpls.2021.640394

The systematic management of scientific and research data, long-term data storage, backup and accessibility, allowing us to network data both nationally, and internationally is increasingly important. The bioinformatics community pushes forward implementing FAIR principles and open science into practice, such as the plant phenomics and genomics repository (Arend et al., 2016).

The various topics discussed during this RT can be summarized into four major categories. (1) **Biodiversity and information systems** includes contributions regarding diversity studies and the embedding of data in information systems. (2) **Distributed computing, tools and infrastructures,** on the other hand, addresses the description, benchmarking and fundamental IT infrastructure of new research software tools and pipelines. (3) **Breeding informatics** is about insights into genome sequence analysis and new challenges in plant breeding. (4) **Image-based analysis and data visualization** presents methods and tools that can be used in the optical and exploratory analysis of plant traits.

CONTRIBUTIONS

The study by Leidenfrost et al. investigates bumble bees as important crop pollinators, examining their food preferences by collecting and sequencing pollen samples. They also compared the results of Illumina short-read technology with Nanopore MinION sequencing. Due to the error-prone nature of Nanopore data, interpretation of these results was more challenging than those of Illumina data. However, the authors were able to conclude that there were fewer errors from the short-read sequencing data, which enabled the discovery of shorter genetic markers in Illumina data in contrast to Nanopore data. This revealed that bumble bees require greater plant diversity than only crops to meet their foraging preferences [relating to the sub-categories of this RT on (1) Biodiversity and information systems and (2) Distributed computing, tools and infrastructures].

Different Illumina marker platforms were used by Soleimani et al. to analyze the effects of core set selection methods in wheat. For this purpose, they introduced a new 15K SNP array, focussing on providing a reliable and cost-effective alternative to other available platforms. They were able to show that the popular *k*-medoids method performs as well as other core selection methods, such as Core Hunter 3 (De Beukelaer et al., 2018) to capture the diversity of a population in a smaller core set [categories (1) Biodiversity and information systems, (2) Distributed computing, tools and infrastructures, and (3) Breeding informatics].

Chu et al. highlight a comparison of different marker systems in bread wheat and the influence on genetic diversity and the prediction ability. While array-based SNP markers showed an ascertainment bias leading to underestimation of diversity within the population, GBS derived markers showed the highest potential as the method of choice for (pre-)breeding programs [categories (1) Biodiversity and information systems and (2) Breeding informatics].

Grehl et al. showed the strengths and weaknesses of different mapping tools for whole-genome bisulfite sequencing in several plant species in both simulated data sets (*Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum*, and *Zea mays*) as well as the real-world data of *Glycine max*. They recommend using BSMAP (Xi and Li, 2009) for its speed and high precision and Bismark (Krueger and Andrews, 2011) for its memory footprint, high precision and the high number of uniquely mapped reads [category (1) Distributed computing, tools, and infrastructures].

Anderson and Murray developed an open-source R function (R/UAS::plotshpcreate) to enable the detection of small plots with remote sensing technologies, such as Unoccupied Aerial Systems (UAS). This allows the creation of multi-polygon shapefiles that also contain information about the experimental design, field orientation and plot dimensions [categories (1) Distributed computing, tools and infrastructures, and (2) Image-based analysis and data visualization].

Lee et al. presented a new genome sequence assembly for winter oilseed rape (*Brassica napus*) accession "Express 617." They used a complex sequencing and assembly strategy with a backbone of 50x Pacific Biosciences long reads, supported by Illumina short-reads, optical map data and genetic maps [relating to categories (1) Biodiversity and information systems and (2) Breeding informatics].

In their article, Santantonio et al. articulate and analyze the potential and challenges of implementing Genomic Selection (GS) in the public breeding programs of developing countries. Proof-of-concept studies were conducted by ICRISAT and CIMMYT in chickpea and maize to examine potential approaches for GS implementation. The authors also discuss the need to develop breeding informatics capabilities to realize large-scale genomic breeding strategies. As an outcome, the authors recommend a multi-phased implementation of GS, (1) building informatics capabilities, optimizing trail design to build cost-effective training sets, (2) increasing selection intensity in the early stage variety development pipeline, early recycling of lines as parents and reducing the number of testing seasons before variety release, and (3) implementation of rapid-cycle recurrent selection to reduce generation intervals toward the biological limits of the species. In this stepwise approach, the genotyping of lines will deliver a series of benefits from the very beginning of the implementation [categories (1) Biodiversity and information systems and (2) Breeding informatics].

Holtgräwe et al. presented a new draft genome assembly of the vitis rootstock "Börner," which is of particular interest for breeders, as this hybrid carries several resistance loci against downy mildew. Using a combination of different short-read sequencing technologies (454 and Illumina) and the incorporation of BAC end sequences, they succeeded in partially separating the resulting contigs into two haplophases. In addition, they generated molecular markers (SNVs and SSRs) and were able to use this new resource to narrow down the position of the resistance locus *Rpv14* (Ochssner et al., 2016) to < 0.5 Mbp on chromosome 5 [categories (1) Biodiversity and information systems and (2) Breeding informatics].

In their study, Narisetti et al. presented an algorithm for segmentation/detection of wheat spikes that relies on a pre-trained neural network classifier. Previous similar approaches

(Qiongyan et al., 2017) applied to images of European wheat cultivars failed to detect spikes growing in the middle of the plant surrounded by multiple leaves of similar color, and textural features. To enhance detection of spikes by suppressing linear leaf contours the Frangi edge filter (Frangi et al., 1998) was applied [categories (1) Distributed computing, tools and infrastructures and (2) Image-based analysis and data visualization].

König et al. implemented a sophisticated web-based visual analysis tool that enabled them to impressively illustrate the high diversity of plant genetic resources of barley species contained in genebanks around the world [categories (1) Biodiversity and information systems, (2) Distributed computing, tools and infrastructures, (3) Breeding informatics and (4) Image-based analysis and data visualization].

CONCLUSIONS

The symbiotic relationship between bioinformatics and plant sciences not only leads to a better understanding of crop biology, it also enriches bioinformatics with powerful methods, theoretical approaches, standards, and software tools. While this RT on Applied Bioinformatics in Crops only covers a few of the recent developments, the overall progress that has been made in recent years is enormous and manifold. These developments range from advances in distributed computing, tools and infrastructures as a backbone for all bioinformatics work, to new or updated information systems mainly in the frame of biodiversity studies. There has been progress in using breeding informatics to facilitate advances in genome sequence analysis, and new methods and tools for image-based analysis and data visualization have proved indispensable for high-throughput phenotyping. The high degree of interdisciplinarity, which becomes apparent when looking at the categories and topics of the contributions for this RT, is worth mentioning and can be seen as evidence of good connectivity within this community. All advances mentioned are crucial to accelerating the development of stress-tolerant elite crops and to improving breeding strategies, both for increased yield and yield stability, as well as rapid breeding cycles. This progress will also depend on the availability of multi-omic and phenomic data sets, particularly training data, to guarantee precise predictions, and the adoption of new bioinformatics technologies. The RT does not cover the topic of systems biology and modeling in crop and plant sciences, which is essential for the progress of the entire research field as we learned during several discussions among conference participants. It is crucial to link the observed genomic and phenotypic variation and to integrate multimodal omic data into coherent frameworks to unravel and understand underlying molecular processes. These integrative models allow for a systems-level understanding and enable research to easily test hypotheses, as well as the effects of the disturbances and perturbations directing field and wet-lab experiments. They pave the way for the development of elite crops and will help breeders to intelligently and rapidly adapt breeding strategies.

AUTHOR CONTRIBUTIONS

M-AB, SB, JS, EG, and US co-wrote this editorial based on the contributions to this Research Topic. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to express our appreciation and thanks to all authors and reviewers who contributed to this Research Topic. We are also grateful to all participants of the Fifteenth *Gatersleben Research Conference on Applied Bioinformatics in Crops*, whose contributions and lively discussions motivated us to edit and compile this Research Topic together with Frontiers. Finally, we would like to thank the financial supporters: Deutsche Forschungsgemeinschaft—DFG (SCHO 1420/4-1), the German Network for Bioinformatics Infrastructure—de.NBI (FKZ 031A536A), KWS Saat SE, SGS-TraitGenetics, ELIXIR Germany, Green Gate Gatersleben, and the IPK Gatersleben, who enabled us to organize the Fifteenth *Gatersleben Research Conference*.

REFERENCES

Arend, D., Junker, A., Scholz, U., Schüler, D., Wylie, J., and Lange, M. (2016). PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* 2016:baw033. doi: 10.1093/database/baw033

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi: 10.1038/s41477-020-0733-0

De Beukelaer, H., Davenport, G. F., and Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* 19:203. doi: 10.1186/s12859-018-2209-z

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). *"Multiscale vessel enhancement filtering," in Medical Image Computing and Computer-Assisted Intervention - MICCAI'98.* Lecture Notes in Computer Science, Vol. 1496, eds W. M. Wells, A. Colchester, and S. Delp (Cambridge MA: Massachusetts Institute of Technology). doi: 10.1007/BFb0056195

Krueger, F., and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572. doi: 10.1093/bioinformatics/btr167

Ochssner, I., Hausmann, L., and Töpfer, R. (2016). "Rpv14, a new genetic source for" Plasmopara viticola resistance conferred by Vitis cinerea. *Vitis J. Grapevine Res.* 55, 79–81. doi: 10.5073/vitis.2016.55.79-81

Qiongyan, L., Cai, J., Berger, B., Okamoto, M., and Miklavcic, S. J. (2017). Detecting spikes of wheat plants using neural networks with Laws texture energy. *Plant Methods* 13:83. doi: 10.1186/s13007-017-0231-1

Soltis, P. S., Nelson, G., Zare, A., and Meineke, E. K. (2020). Plants meet machines: prospects in machine learning for plant biology. *Appl. Plant Sci.* 8:e11371. doi: 10.1002/aps3.11371

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10:214. doi: 10.3389/fgene.2019.00214

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 10:232. doi: 10.1186/1471-2105-10-232

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 Plant Science

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fpls.2020.00042

Suitability of Single-Nucleotide Polymorphism Arrays Versus Genotyping-By-Sequencing for Genebank Genomics in Wheat

*Jianting Chu[1], Yusheng Zhao[1], Sebastian Beier[1], Albert W. Schulthess[1], Nils Stein[2], Norman Philipp[1], Marion S. Röder[1] and Jochen C. Reif[1,3]**

[1] Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [2] Department of Genebank, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [3] Faculty of Sciences III - Agricultural and Nutritional Sciences, Earth Sciences and Computer Science, Martin-Luther-University Halle-Wittenberg, Halle/Saale, Germany

Genebank genomics promises to unlock valuable diversity for plant breeding but first, one key question is which marker system is most suitable to fingerprint entire genebank collections. Using wheat as model species, we tested for the presence of an ascertainment bias and investigated its impact on estimates of genetic diversity and prediction ability obtained using three marker platforms: simple sequence repeat (SSR), genotyping-by-sequencing (GBS), and array-based SNP markers. We used a panel of 378 winter wheat genotypes including 190 elite lines and 188 plant genetic resources (PGR), which were phenotyped in multi-environmental trials for grain yield and plant height. We observed an ascertainment bias for the array-based SNP markers, which led to an underestimation of the molecular diversity within the population of PGR. In contrast, the marker system played only a minor role for the overall picture of the population structure and precision of genome-wide predictions. Interestingly, we found that rare markers contributed substantially to the prediction ability. This combined with the expectation that valuable novel diversity is most likely rare suggests that markers with minor allele frequency deserve careful consideration in the design of a pre-breeding program.

Keywords: single-nucleotide polymorphism (SNPs), genotyping-by-sequencing (GBS), simple sequence repeats (SSR), genebank genomics, molecular diversity, genome-wide prediction, wheat

OPEN ACCESS

Edited by:
Fabio Marroni,
University of Udine, Italy

Reviewed by:
Davoud Torkamaneh,
University of Guelph, Canada
Francois Belzile,
Laval University, Canada

***Correspondence:**
Jochen C. Reif
reif@ipk-gatersleben.de

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *15 October 2019*
Accepted: *13 January 2020*
Published: *14 February 2020*

Citation:
Chu J, Zhao Y, Beier S, Schulthess AW, Stein N, Philipp N, Röder MS and Reif JC (2020) Suitability of Single-Nucleotide Polymorphism Arrays Versus Genotyping-By-Sequencing for Genebank Genomics in Wheat. Front. Plant Sci. 11:42. doi: 10.3389/fpls.2020.00042

INTRODUCTION

Global agricultural production must be increased by 60% compared to 2005–2007 levels in order to supply an estimated world population of 9 billion in 2050 (Ray et al., 2013; FAO, 2017). The annual yield increases for the four main crops (wheat, corn, rice, and soybean) are about 0.9%–1.6%, which is far below the required one (Ray et al., 2013). It is becoming increasingly difficult to meet this rising global demand as arable land and water become scarcer, average living standards rise, and

investments to increase agricultural productivity grow slowly (Fischer et al., 2014; Laidig et al., 2014). Wheat breeding is a viable and sustainable solution for increasing grain yield and improving yield stability (Borlaug, 1968; Voss-Fels et al., 2019).

The success of wheat breeding strongly depends on the availability of a valuable diversity within breeding populations (Jannink et al., 2010; Rufo et al., 2019). The effective population size in European wheat breeding populations is small with an estimated value of ~30 individuals (He et al., 2017). Therefore, the extension of the genetic diversity of elite wheat breeding pools through the introgression of valuable variation is crucial for increasing the grain yield potential. Moreover, the systematic genotyping of collections was proposed as a first step toward developing new ways and approaches to unlock wheat genetic resources for breeding (Mascher et al., 2019). Genotyping of plant genetic resources (PGRs) was performed for some important crops such as barley (Milner et al., 2019), maize (Romay et al., 2013), and rice (Wang et al., 2018). As far as wheat is concerned, many efforts have focused on how genomic technologies can be used to genotype PGRs (Rasheed et al., 2018). For example, the global landrace collection "Watkins" was genotyped with 41 simple sequence repeat (SSR) markers for 826 landraces from 32 countries (Wingen et al., 2014). A collection of 295 accessions including 136 landraces from 25 countries from the Australian Grains genebank was fingerprinted by genotyping-by-sequencing (GBS) and Diversity Arrays Technology (DArT-seq) (Riaz et al., 2017). An 820k Axiom single-nucleotide polymorphism (SNP) array as well as a 35k subset were developed by genotyping 43 bread wheat lines including their wild accessories (Winfield et al., 2016; King et al., 2017). The GBS platform was also used for genotyping "Creole" landraces conserved in CIMMYT's genebank (Vikram et al., 2016), a sample of 62 diverse wheat lines including 26 landraces (Jordan et al., 2015), a set of 1,143 accessions of *Aegilops tauschii* (Singh et al., 2019) and a set of 1,423 spring bread wheat germplasm including 561 landrace accessions (Sehgal et al., 2015). These recent works present the potential of introducing exotic alleles present in these PGRs to improve elite wheat lines. In this sense, the genomic data not only allow to estimate the neutral molecular diversity of genetic resources as compared to that of elite lines (He et al., 2019) but also to combine it with phenotypic information in order to find novel valuable functional genetic variation, i.e. genes/alleles/haplotypes (e.g., Milner et al., 2019) or to build up genome-wide prediction models to select promising candidates for (pre)breeding (Yu et al., 2016). Whole-genome sequencing of entire collections is currently not affordable in large-genome species such as wheat and therefore attempts have been mainly focused on cost-effective genotyping platforms (Milner et al., 2019). Several marker platforms have been developed in wheat in the past (Elbasyoni et al., 2018). SSR markers (Röder et al., 1995; Röder et al., 1998) were replaced by diversity array technology (DArT markers; Wenzl et al., 2004), GBS (Elshire et al., 2011; Poland et al., 2012), and array platforms for scoring SNPs (Cavanagh et al., 2013; Wang et al., 2014; Winfield et al., 2016). The disadvantage of most cost-efficient genotyping platforms in contrast to whole-genome sequencing is that an ascertainment bias can be introduced by designing the marker platforms using a limited set of individuals (Clark et al., 2005). This has been described for instance in maize (e.g., Frascaroli et al., 2013). An ascertainment bias can impact the estimates of the diversity within populations but seems to be of minor relevance for the estimates of the overall population structure (Heslot et al., 2013; Alipour et al., 2017; Eltaher et al., 2018; Bhatta et al., 2018) or further downstream applications such as genome-wide predictions (Heslot et al., 2013; Jiang et al., 2015; Elbasyoni et al., 2018). For wheat, only a few studies have compared the accuracy of genome-wide prediction between SSR and SNP array markers (e.g., Jiang et al., 2015), between GBS and DArT markers (e.g., Heslot et al., 2013), and between GBS and SNP array markers (e.g., Elbasyoni et al., 2018). The results heavily depend on the underlying germplasm, while studies on the relevance of an ascertainment bias on diversity estimates and genome-wide predictions in wheat genetic resources are rare. Furthermore, it is also promising to test whether genetic information from different marker platforms is complementary and whether their integrated use can boost prediction accuracies.

The objectives of our study were to 1) compare the relevance of an ascertainment bias on the genetic diversity estimated by SSR, GBS, and SNP array markers in a wheat population comprising PGRs and European elite lines, 2) contrast the prediction ability obtained using the three marker platforms, and 3) investigate the potential and limits of genome-wide prediction models exploiting the complementarity of different marker platforms.

MATERIALS AND METHODS

Genotyping and Population Genetic Analyses

We fingerprinted 378 winter wheat (*Triticum aestivum* L.) genotypes: 190 lines represent the elite breeding pool exploited in Europe (Elite) and 188 genotypes represent a random sample of PGRs maintained at the genebank of the IPK Gatersleben, Germany. Details on the plant material have already been published (Philipp et al., 2018). The 378 wheat lines were characterized using (1) an Infinium 90,000 SNP array for 174 genotypes out of 571 samples (Wang et al., 2014) and a derived Infinium 15,000 SNP array for 204 genotypes out of 782 samples (Boeven et al., 2019), (2) GBS (Wendler et al., 2014), and (3) 19 SSR markers (Plaschke et al., 1995; Röder et al., 1995; Röder et al., 1998). The 90,000 SNP array data were used from a previously published study (Zanke et al., 2014a; Zanke et al., 2014b; Zanke et al., 2015). The development of the 15,000 SNP array and genotyping was performed by TraitGenetics GmbH (www.traitgenetics.com) and the SNPs represent a subset of markers from the 90,000 SNP array (Wang et al., 2014). The GBS data were generated and processed following established protocols (Himmelbach et al., 2014; Wendler et al., 2014). Briefly, digestion of genomic DNA was done with the enzymes PstI and MspI (New England Biolabs). Up to 190 individually

barcoded samples were pooled per lane equimolarly and sequenced on the Illumina HiSeq 2000 device with 1 x 107 cycles in single-end mode using custom sequencing primer (Meyer and Kircher, 2010) according to the manufacturer's instructions. In total, five lanes of a single flow cell were sequenced with an average output of 3,052,589 raw reads per sample (ranging from 322,285 to 10,758,745 reads per sample) for 378 individuals (**Supplementary Table 1**). Following adapter trimming with cutadapt (Martin, 2011), reads were mapped to the reference genome sequence of bread wheat cultivar Chinese Spring (IWGSC, 2014) with BWA-MEM version 0.7.13 (r1126) (Li, 2013) using the -M option to mark shorter split hits as secondary. Mappings were transformed into the BAM format with SAMtools version 1.3 (Li et al, 2009). Novosort version 3.02.12 [1] was applied to sort and index records by position. BAM files were merged by genotype with Picard[2]. We called variants using the SAMtools/BCFtools pipeline version 1.3 (Li et al, 2009) with mpileup parameter set to "-DV". A custom awk script was applied for initial filtering of genotype calls in the following manner: Bi-allelic sites with a minimum mapping quality score of 40 were called for homozygous and heterozygous genotype calls that were supported by at least two and four reads, respectively. We coded the SNP array and GBS marker data as (0, 1, 2, NA), where 0 and 2 represent the homozygous state for the first and second allele at a particular SNP locus, respectively, 1 represents the heterozygote class, and NA refers to missing values. As to multi-allelic SSR markers, if the allele appears for a certain genotype, it was coded as 1, if not, then 0. After that, this coding was also used for SSR markers assuming that each allele is a marker. We assessed the quality of the marker data in two steps: firstly, we deleted markers showing more than 5% of missing values, and then, we excluded the monomorphic markers [allele frequency (AF) = 0 or = 1]. After the quality assessment, 12,490 SNP array markers, 31,230 GBS markers, and 170 SSR alleles remained in the matrix. We then explored the genetic diversity based on these filtered markers without imputation and imputed the missing values according to the distribution of allele frequency for genomic prediction.

In order to compare properties between Elite and PGRs for each marker dataset, we calculated the minor allele frequency (MAF), population heterozygosity (H), and polymorphism information content (PIC). The standard deviations (SD) of these parameters were derived by means of bootstrapping with 1,000 rounds. We evaluated the genetic diversity from each group and calculated the Rogers' distances (RD) between pairs of genotypes. SDs were obtained by resampling genotypes without replacement with 1,000 rounds. Principal coordinates analysis (PCoA, Gower, 1966) was performed to investigate the population structure. PCoA was implemented with the function "cmdscale" from the R package "stats" [3]. The relatedness of each pair of marker datasets was assessed through the Mantel correlation of their corresponding RD matrices (Mantel, 1967). Detailed information on the implementation of the population genetic analyses is outlined in the **Supplementary Material**.

Field Trials and Phenotypic Data Analysis

For 339 genotypes (188 Elite and 151 PGR), phenotypic data were available. The 339 genotypes (or subsets) were phenotyped for grain yield (GY) (Mg ha^{-1}) and heading date (HD) (days since 1 January) in three field experiments (**Table 1**). Experiment 1 comprised field trials of up to 278 genotypes evaluated in Gatersleben, Germany, and Malchow, Germany. The trials were performed in the year 2015 following an alpha-lattice design with two replicates (for details, see Philipp et al., 2018). Plot sizes were 5 m^2 in Gatersleben and 3.75 m^2 in Malchow. Experiment 2 included 166 out of the 188 elite lines and further 164 varieties (for details, see Zanke et al., 2014b; Kollers et al., 2013; Schulthess et al., 2017). Briefly, the experimental design was an alpha design with two replicates. The field trials were conducted in five locations during years 2009 and 2010, giving rise to eight location × year combinations (environments). Plot sizes ranged from 5 to 6.75 m^2. Experiment 3 comprised field evaluation at five locations during 2016 and included 12 out of the 188 elite lines and 61 out of the 151 PGR. Briefly, the experimental design was an unreplicated alpha design (for details, see Boeven et al., 2019). Plot sizes ranged from 7.56 to 12 m^2. Across the three experiments, the 188 elite lines and the 151 PGR were evaluated in up to 15 environments for grain yield and in up to 11 environments for HD, respectively.

We performed outlier tests and implemented a Bonferroni-Holm test standardized by the re-scaled median absolute deviation (MAD) (BH–MADR) at a significance level ($P < 0.05$) (Bernal-Vasquez et al., 2016). Thereafter, best linear unbiased estimations (BLUEs) and heritability for GY and HD were independently obtained using a two-stage approach.

TABLE 1 | Description of the environments used for evaluating grain yield and heading date (HD).

Experiment	Location	Year	No. of Elite	No. of PGR	Grain yield	Heading date (HD)
1	Gatersleben	2015	187	91	×	
	Malchow	2015	186 (184)*	91	×	×
2	Andelu	2009	166	0	×	×
	Andelu	2010	166	0	×	×
	Janville	2010	166	0	×	×
	Saultain	2010	166	0	×	×
	Seligenstadt	2009	166	0	×	×
	Seligenstadt	2010	166	0	×	×
	Wohlde	2009	166	0	×	×
	Wohlde	2010	166	0	×	×
3	Hohenheim	2016	12	61	×	×
	Renningen	2016	12	61	×	×
	Gatersleben	2016	12	61	×	
	Schackstedt	2016	12	61	×	
	Böhnshausen	2016	12	61	×	

** The number of elite lines for Malchow (2015) are different between grain yield (186) and HD (184).*

[1] www.novocraft.com/documentation/novosort-2/
[2] https://broadinstitute.github.io/picard/
[3] https://cran.r-project.org/web/packages/STAT/STAT.pdf

First, BLUEs of each genotype within each single environment were estimated by fitting the following model:

$$P = 1_n\mu + G + R + B + e \quad (1)$$

in which, $\boldsymbol{P}$ contains the phenotypic values of GY or HD for each plot, $\boldsymbol{\mu}$ corresponds to the overall mean, $\boldsymbol{G}$ represents the genotype effect, $\boldsymbol{R}$ stands for the effect of the replication, $\boldsymbol{B}$ is the effect of incomplete blocks, and $\boldsymbol{e}$ refers to the error term of the model. In the model, only $\boldsymbol{\mu}$ and $\boldsymbol{G}$ were treated as fixed effect, while all other components were assumed to be random effects.

Second, the BLUEs of genotypes across all environments were estimated fitting the following model:

$$Y = 1_n\mu + G + E + G\times E + e \quad (2)$$

in which, $\boldsymbol{Y}$ contains the genotypic effects estimated within each environment using Equation (1), $\boldsymbol{\mu}$ is the fixed effect of the overall mean, $\boldsymbol{G}$ corresponds to the fixed effects of genotypes across environments, $\boldsymbol{E}$ stands for random environment effects, $G\times E$ indicates the random effects of interaction between genotype and environment, and $\boldsymbol{e}$ is a random error term. Equations (1) and (2) were fitted using the mixed model R package ASReml-R (Butler et al., 2009).

Model (2) was also used to estimate the variances and heritability of each trait. During the computation for variances and heritability, $\boldsymbol{\mu}$ is taken as fixed effect, while all other components in the model are assumed as random. Thereby, we calculated the broad-sense heritability ($\boldsymbol{H}^2$) as:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{G\times E}^2/n + \sigma_e^2/(r\times\bar{n})} \quad (3)$$

in which, $\boldsymbol{\sigma_G^2}$ is the variance of genotypes, $\boldsymbol{\sigma_{G\times E}^2}$ indicates the variance of genotype times environment interaction, $\boldsymbol{\sigma_e^2}$ stands for the variance of error terms, $\bar{\boldsymbol{n}}$ is the average number of environments in which genotypes were evaluated, and $\boldsymbol{r}$ represents the average number of replications.

Genome-Wide Prediction

A genomic best linear unbiased prediction (GBLUP) model was implemented, with the co-variance matrix (G matrix) derived from SNP array, GBS, or SSR marker datasets. We employed single G matrix (single-kernel) or their combination (multi-kernel). The GBLUP model of the multi-kernel model was:

$$Y = 1_n\mu + g_{SNP} + g_{GBS} + g_{SSR} + e \quad (4)$$

Where $\boldsymbol{Y}$ contains the BLUEs for each trait, $\boldsymbol{g_{SNP}}$, $\boldsymbol{g_{GBS}}$ and $\boldsymbol{g_{SSR}}$ are random genetic effects derived from different markers, with $\boldsymbol{g_{SNP} \sim N(0, A_{SNP}\sigma_{G_1}^2)}$, $\boldsymbol{g_{GBS} \sim N(0, A_{GBS}\sigma_{G_2}^2)}$, $\boldsymbol{g_{SSR} \sim N(0, A_{SSR}\sigma_{G_3}^2)}$, and $\boldsymbol{e \sim N(0, I\sigma_e^2)}$, while $\boldsymbol{A_{SNP}}$, $\boldsymbol{A_{GBS}}$ and $\boldsymbol{A_{SSR}}$ are the numerator relationship matrix calculated using SNP array, GBS, or SSR marker datasets, respectively, according to VanRaden (2008) and $\boldsymbol{\sigma_{G_1}^2}$ to $\boldsymbol{\sigma_{G_3}^2}$ are the respective genetic variances of each component of the model. For single-kernel models, we used the $\boldsymbol{g_{SNP}}$, $\boldsymbol{g_{GBS}}$, and $\boldsymbol{g_{SSR}}$ individually. The implementation of the models is described in detail in the **Supplementary Material**.

We applied a random resampling method for fivefold cross validation to investigate the prediction ability. In each cross validation, the population was divided into a training (80%) and a test set (20%). We used the training set to build the mixed model function, which was then used to predict the genetic value of the test set. The prediction ability was calculated as the Pearson correlation between estimated genetic values and the observed values in the test set. We performed 1,000 rounds of cross validation and recorded the mean and SD for these 1,000 correlation coefficients. The genomic prediction model was fitted using the "BGLR" R-package (Pérez and de los Campos, 2014). Besides GBS data generation, all computational methods were implemented in R environment (R 3.4.3, R Core Team, 2018).

RESULTS

Molecular Diversity Estimated From SNP Array, GBS, and SSR Marker Data

We found for the SNP array markers ~5–6 times higher estimates of MAF, H, PIC, and RD than for the GBS markers considering the total population of 378 lines (**Table 2**; **Supplementary Figures 1** and **2**). In contrast, the values of H, PIC, and RD for the SNP array markers were only half as large as for the SSR markers, however, MAF for SNP array markers are roughly two times larger than for the SSR markers. Moreover, the mean values of these indices within the sample of 190 elite lines were generally lower compared to the population of PGR, regardless of the marker system. This shows the large molecular diversity of wheat accessions hosted at the genebank of the IPK Gatersleben.

The SNP array markers followed a uniform pattern of MAF ranging from 0 to 0.5 (**Figure 1**), especially for the PGR population. In contrast, GBS markers were characterized by very low MAF in the range between 0 and 0.05. This suggests that GBS markers are more reliable in detecting the profile of rare alleles compared to SNP array markers. The distribution of MAF from SSR was derived from only 19 markers, and therefore the index spectra were quite sparse, which has to be considered when

TABLE 2 | The mean and standard deviations (SD) of minor allele frequency (MAF), population heterozygosity (H), polymorphism information content (PIC), and average Rogers' distances (RD) for SNP array (SNP), genotyping-by-sequencing (GBS), and SSR markers.

Index	Marker set	All genotypes		Elite lines		PGRs (plant genetic resources)	
		Mean	SD	Mean	SD	Mean	SD
MAF	SNP	0.2438	0.0023	0.2172	0.0029	0.2480	0.0034
	GBS	0.0439	0.0006	0.0382	0.0005	0.0463	0.0009
	SSR	0.1382	0.0004	0.1381	0.0006	0.1385	0.0006
H	SNP	0.3299	0.0027	0.2961	0.0035	0.3336	0.0038
	GBS	0.0662	0.0009	0.0571	0.0008	0.0702	0.0015
	SSR	0.6765	0.0059	0.6286	0.0082	0.6924	0.0081
PIC	SNP	0.2418	0.0019	0.2177	0.0025	0.2443	0.0027
	GBS	0.0525	0.0008	0.0448	0.0006	0.0555	0.0012
	SSR	0.6449	0.0064	0.5930	0.0084	0.6655	0.0087
RD	SNP	0.3312	0.0528	0.2987	0.0472	0.3368	0.0532
	GBS	0.0651	0.0143	0.0561	0.0094	0.0696	0.0155
	SSR	0.6880	0.1190	0.6482	0.1186	0.7045	0.1190

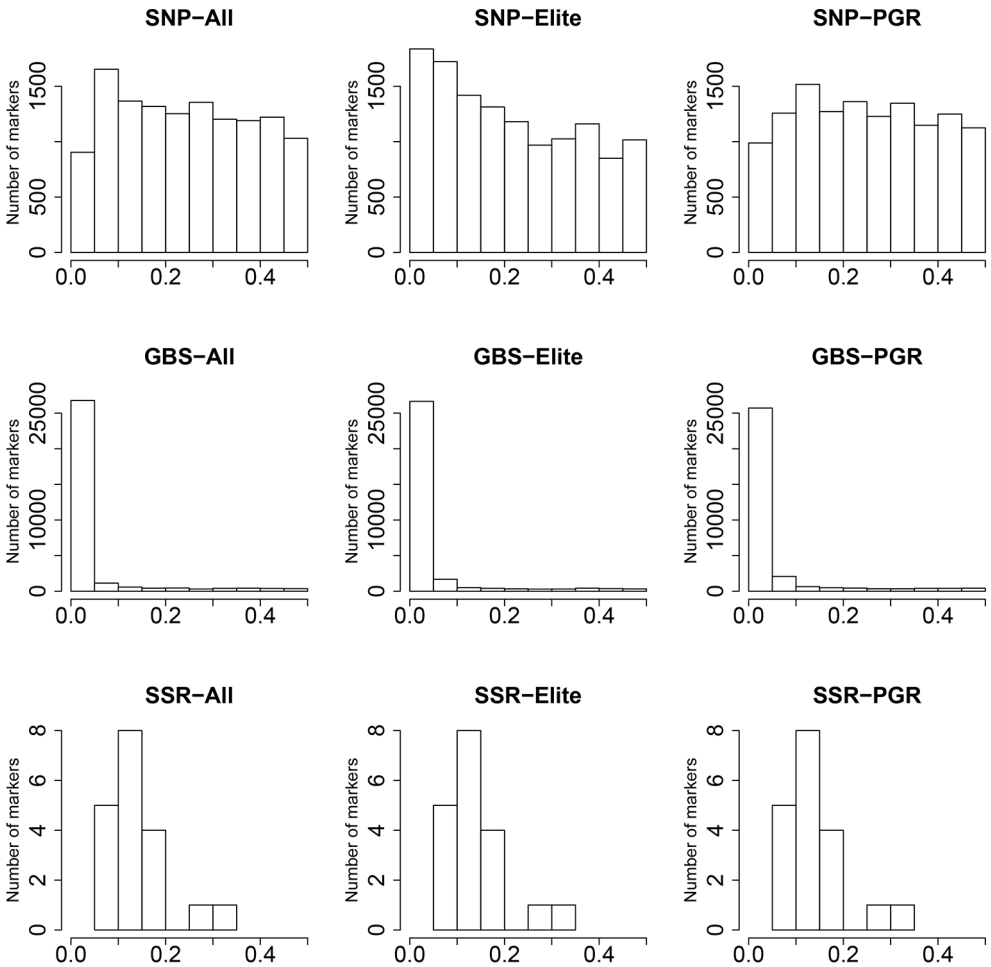

FIGURE 1 | Distribution of minor allele frequencies (MAF) (x-axis) for single-nucleotide polymorphism (SNP) array, genotyping-by-sequencing (GBS), and SSR markers. Results are shown for the total population (All), the elite lines (Elite), and the plant genetic resources (PGR).

interpreting the results. In this context, we observed a peak at the MAF range between 0.05 and 0.2 for SSR markers.

The picture of the relatedness among the lines estimated on the basis of SNP array or GBS markers was similar (**Supplementary Figure 3**) and the correlation between distance matrices was up to r = 0.83 for the PGR population (**Table 3**). The correlations were significantly lower between SSR- and SNP array-based distance matrices with maximum r values of 0.48 and 0.52 when comparing SSR- with GBS-based and SNP array-based distance matrices, with both maximum values observed again in PGR.

TABLE 3 | Correlation between Rogers' distance (RD) matrixes calculated using data from SNP array (SNP), genotyping-by-sequencing (GBS), and SSR markers.

	All	Elite	PGR
SNP—GBS	0.818	0.683	0.830
GBS—SSR	0.454	0.414	0.476
SNP—SSR	0.500	0.442	0.520

Results are shown for the total population (All), the elite lines (Elite), and the plant genetic resources (PGR). Correlations were significantly (P < 0.001) larger than zero according to a Mantel test.

The first, second, and third principal coordinates (PC1, PC2, and PC3) calculated based on the SNP-array data explained 10.42%, 4.62%, and 2.95% of the molecular variation, respectively (**Figure 2**, **Supplementary Table 2**). Elite lines and PGR were separated with respect to PC1. The distribution along PC2 and PC3 reflected the diversity within elite lines and PGR. A similar pattern was observed for the principle coordinate analysis based on the GBS data: Elite lines were separated from PGR with respect to PC1 and diversity within subpopulations was represented mainly by PC2 and PC3. The molecular variance explained by PC1, PC2, and PC3 was lower for the GBS compared to the SNP array data and amounted to 5.73%, 2.31%, and 1.74%, respectively. Similarly, the range of PC for the GBS marker was about 1/10 times of that of the SNP array data (**Figure 2**, **Supplementary Table 2**). For the SSR data, the differentiation between elite lines and PGR was less pronounced.

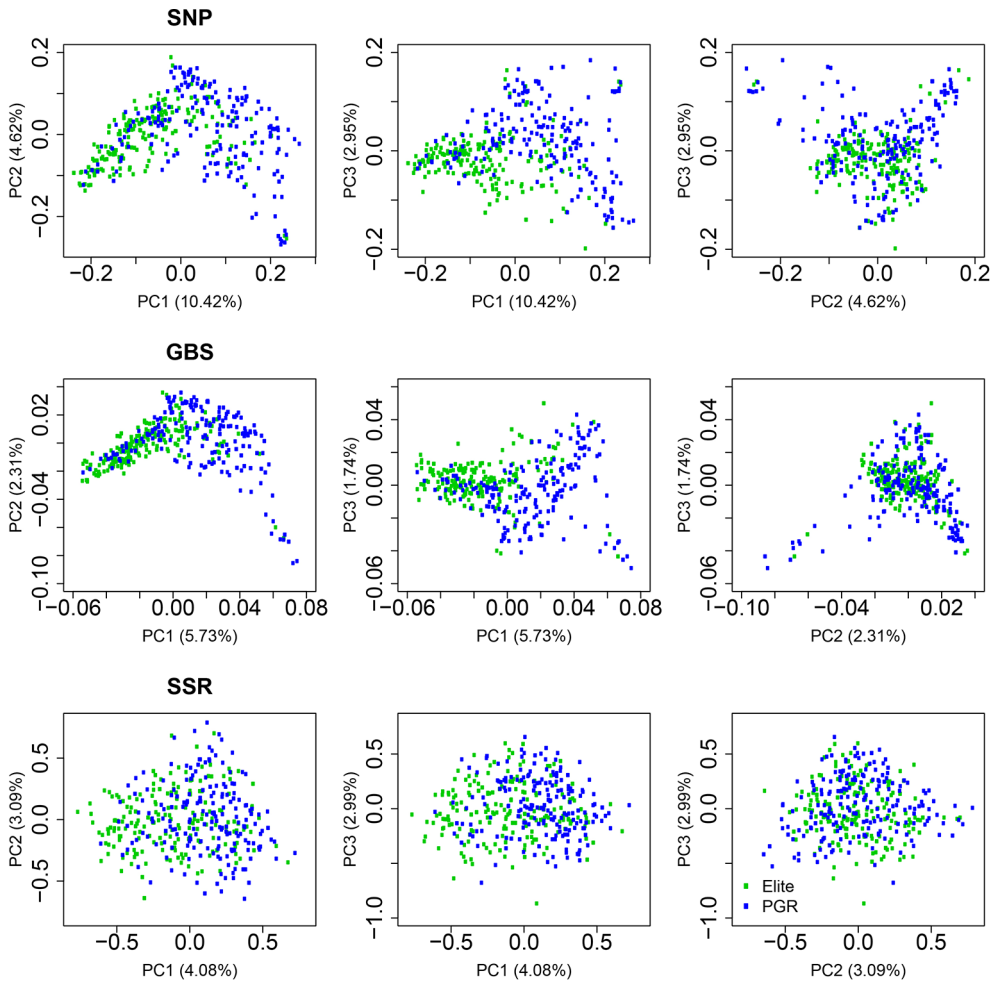

FIGURE 2 | Principal coordinate analyses using data from single-nucleotide polymorphism (SNP) array, genotyping-by-sequencing (GBS), and SSR markers. Results are shown for the total population (All), the elite lines (Elite), and the plant genetic resources (PGR). PC1, PC2, and PC3 refer to the first, second, and third principal coordinate, respectively. Explained proportion of molecular variation is given in brackets.

In this case, PC1, PC2, and PC3 accounted for 4.08%, 3.09%, and 2.99% of the molecular variation.

Comparison and Application of SNP Array, GBS, and SSR Markers in Genome-Wide Prediction

We estimated BLUEs of grain yield and HD for 339 of the 378 fingerprinted genotypes, including 188 Elite lines and 151 PGR. The BLUEs approached a bell-shaped distribution for both traits (**Supplementary Figure 4**). Heritability was 0.94 and 0.98 for grain yield and HD, respectively, which illustrates the high quality of the phenotypic data.

The phenotypic data were combined with the different marker datasets and the prediction abilities for the combination of the different marker kernels in the total population of 339 lines were evaluated. We observed comparable prediction abilities for grain yield for the GBS and SNP array data, amounting to an average of 0.829 (**Figure 3**). The same picture was observed when comparing the prediction abilities for HD, but with a slightly lower level (0.741 and 0.710 for SNP array and GBS marker data, respectively). In contrast, the prediction abilities of SSR markers for grain yield (0.633) and HD (0.571) were significantly lower compared to SNP array and GBS markers. For grain yield, the prediction ability of the two-kernel model from the combination of SNP array and GBS markers (S-G) was slightly higher than that of the combination of GBS and SSR (G-S), followed by the combination of SNP array and SSR markers (S-S) (**Figure 3**). The highest prediction ability was achieved for the three-kernel model of the combination of SNP array, GBS, and SSR markers (S-G-S) (**Figure 3**). All in all, prediction abilities of the different kernel models were comparable with the only exception being the single model based on the G matrix derived from SSR markers. For the HD, the trends in prediction abilities of the different models were similar, but with lower values.

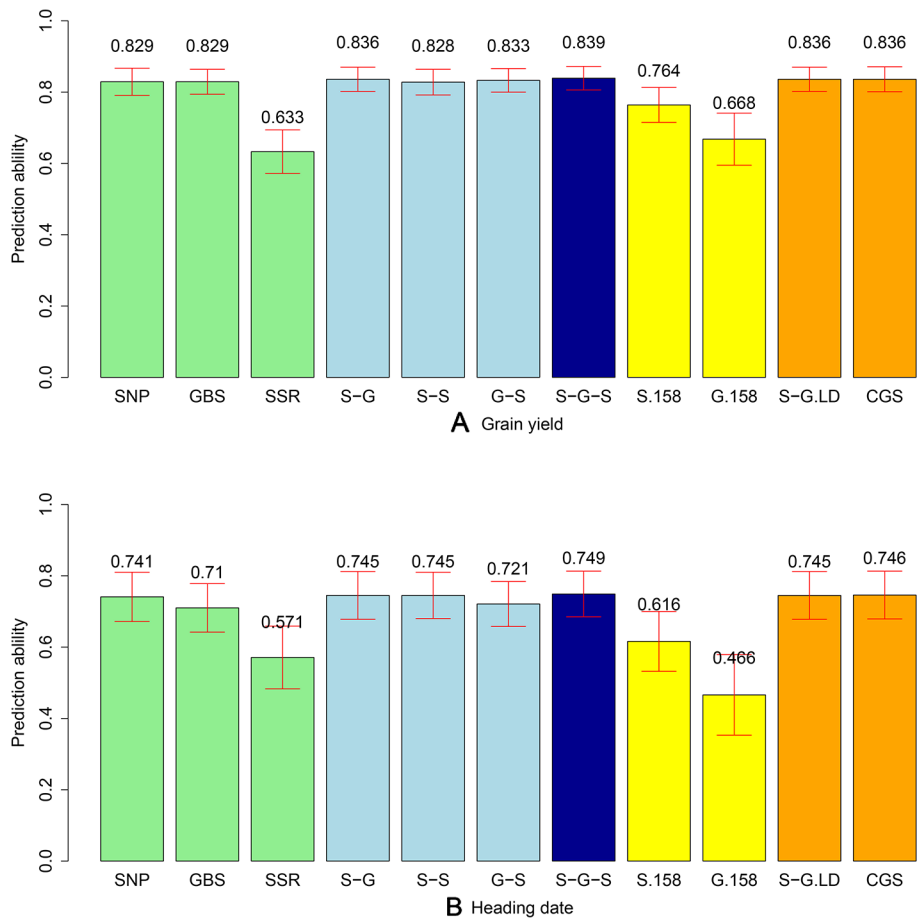

FIGURE 3 | Bar plot of average prediction abilities derived from 1,000 cross-validations from different prediction models for **(A)** grain yield and **(B)** heading date (HD). Single kernel models (green) were used for data from single-nucleotide polymorphism (SNP) array, genotyping-by-sequencing (GBS), and SSR markers. Double-kernel models (light blue) were used combining SNP array and GBS markers (S-G), SNP array and SSR markers (S-S), as well as GBS and SSR markers (G-S). The three-kernel model (dark blue) combined SNP array, GBS, and SSR markers (S-G-S). Subsets of 158 markers from SNP array markers (S.158) and GBS markers (G.158) were used to run the single kernel models (yellow). Moreover, after ignoring the GBS markers with higher linkage equilibrium with SNP array markers, a double-kernel model combing SNP array and remained GBS markers (S-G.LD) and a single-kernel model of the combination of SNP array and remained GBS markers (CGS) (orange) were used. The corresponding standard deviations are illustrated as red bars.

To discard the influence of marker density, we randomly selected 158 SNP array (S.158) or GBS markers (G.158), calculated the G matrices, and evaluated prediction abilities of single-kernel models applying cross validations. In general, the prediction ability of S.158 and G.158 was up to 34.4% lower than the total marker set (**Figure 3**). Interestingly, we observed lower prediction ability with the SSR compared to the S.158 and G.158 panels with the exception of the G.158 prediction for HD. In addition, the decrease in prediction ability was much more pronounced for the G.158 than for the S.158, suggesting an influence of the allele frequency distribution. We further inspected therefore the total set of GBS markers and tested the decrease in prediction abilities for GBS markers in dependence with MAF. The prediction ability decreased for both traits, grain yield and HD, with increasing thresholds of MAF (**Figure 4**). The number of markers decreased mostly in the interval between MAF $0 < 0.05$. Thus, markers with very low MAF contributed substantially to the prediction ability for both traits, suggesting that they are actually important for genome-wide prediction.

Linkage disequilibrium (LD) between markers can impact the prediction ability for the multi-kernel models. We calculated therefore the LD between each pair of SNP array and GBS markers across the 339 lines and deleted the corresponding GBS markers if their LD was higher than $r^2 = 0.95$. After removing 2,826 (9.5%) GBS markers, which were in tight LD, we combined SNP array and remaining GBS markers to build a new dataset (CGS). We then did two *in-silico* experiments: first, we used the double-kernel model based on the SNP array and the GBS data excluding the linked markers (S.G.LD); second, we applied a single-kernel model for CGS. We observed for both traits that the performance of these two models was very close to that of S-G (**Figure 3**). Thus, the influence of linked markers was

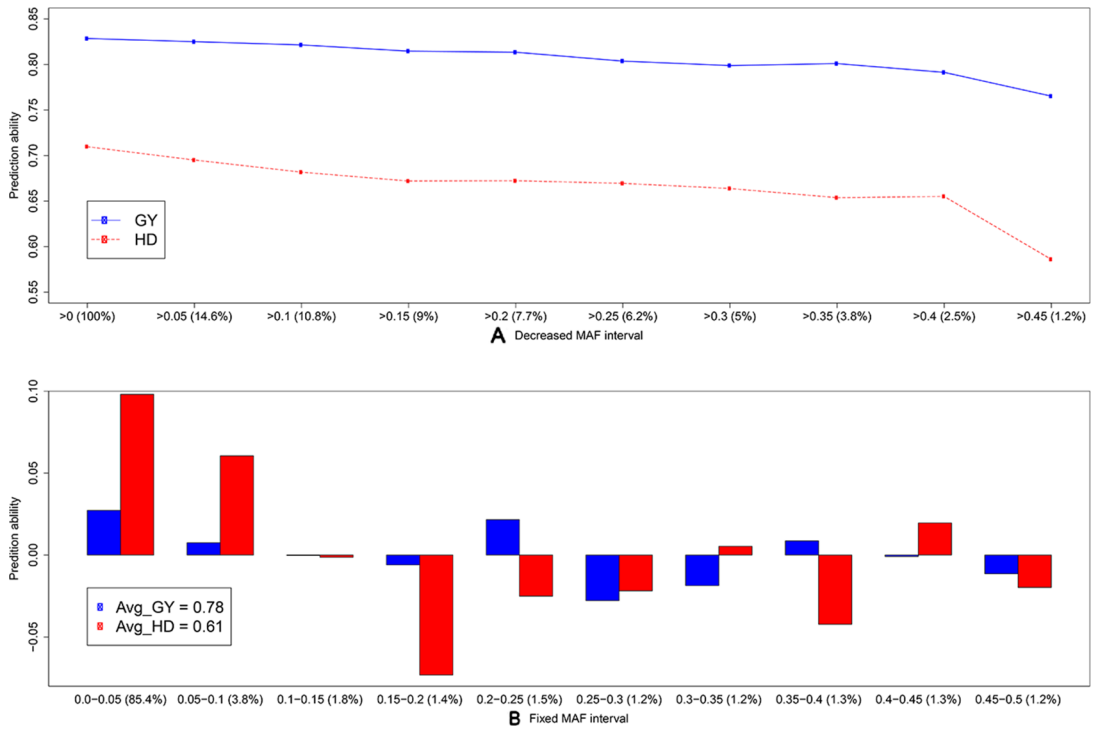

FIGURE 4 | Average prediction ability derived from 1,000 cross-validation using the single-kernel model with a kernel matrix from genotyping-by-sequencing (GBS) markers for grain yield (GY, blue) and heading date (HD, red) for **(A)** decreased minor allele frequency (MAF) interval and **(B)** fixed MAF interval. The percentage (from the total number) of markers within frequency intervals is indicated within brackets. In **(B)**, bars indicate the average differences in prediction ability, and the average prediction abilities are indicated in the legend.

ignorable; however, if a huge number of markers are available, these results also indicate that the computational load can be decreased if linked markers are removed.

DISCUSSION

Data from GBS is typically characterized by a significant proportion of missing values (Elshire et al., 2011). We used a robust strategy to confront the challenges of dealing with missing values and, in a first step, filtered reliable SNPs with less than 5% of missing values. Then we imputed the missing values according to the original distribution of allele frequency for the implementation of genomic prediction. Nevertheless, it has already been shown that increasing the marker density beyond 3,000 SNPs in wheat populations of the size used in our study does not increase the genome-wide prediction ability nor does affect significantly the estimates of the relatedness among accessions (Liu et al., 2016). This is not the case for genome-wide association mapping studies, for which imputing missing values and increasing the marker density boosts the power of QTL detection (e.g., He et al., 2015; Negro et al., 2019). We would like to note that association mapping, however, was not the target of our study.

Genotyping-By-Sequencing Enables Unbiased Estimates of the Genetic Diversity in Wheat Populations

Entire genebank collections have been fingerprinted using different marker technologies (e.g., Romay et al., 2013; Wang et al., 2018; Milner et al., 2019; Singh et al., 2019). In order to limit the costs, the sequence variation being represented is usually reduced. SSR markers, array-based scoring of SNPs, and GBS differ dramatically in the way sequence variation is reduced: GBS depends on the restriction enzymes used (Elshire et al., 2011), while SSR markers and also SNPs from arrays are selected using a subpopulation with limited size (Frascaroli et al., 2013). The 90k SNP array in wheat (Wang et al., 2014), for instance, was developed using data resulting from sequence information of 19 bread wheat and 18 tetraploid lines, as well as previous sequence information on 24 (M Ganal unpublished data; for details see Wang et al., 2014), 23 (Allen et al., 2011), 28 (Cavanagh et al., 2013), and 8 (Pont et al., 2013) wheat genotypes. The panel was selected to cover the global wheat diversity and included several elite wheat lines. The limited number of individuals used for SNP array discovery and the array design can led to a distorted picture of the molecular diversity denoted as ascertainment bias (Clark et al., 2005). Signs of an ascertainment bias are that rare alleles are missed,

polymorphic markers have a high frequency of major alleles and genetic diversity is underestimated in the non-ascertained population (Clark et al., 2005). As already mentioned, H, PIC, and RD absolute estimates were ~5–6 times higher when computed from array-based SNPs than those obtained from GBS data (**Table 2**). Nevertheless, these results must be carefully interpreted, because this observation can be simply caused by a scale issue. In fact, we observed 23%, 24%, and 24% higher values based on H, PIC, and RD within the PGR population compared to elite lines as revealed by GBS, but this increased diversity amounted to only 13%, 12%, and 13% according to SNP array results, respectively. Moreover, for the SNP array data, the number of rare alleles was lower in the PGR population compared to elite lines (**Figure 1**). This was not the case for SNPs resulting from GBS data. Although it is true that the amount of SSR markers is substantially lower when compared to SNP array and GBS markers, which is mainly due to the high cost per data point of SSR markers, SSR markers are still being used by many researchers to study the genetic diversity existent in important crop species like potato (Wang et al., 2019), wheat (Sajjad et al, 2018), and maize (Adu et al, 2019). Moreover, it is interesting to observe that SSR markers are much capable to catch and portray the genetic diversity even with such a low number (19 markers and altogether 170 alleles). Altogether, these findings point to an underestimation of the diversity within the population of PGR versus the set of elite lines using the 90k SNP array, which can be explained by a large proportion of elite lines used to design the 90k SNP array.

The principal coordinate analyses revealed a comparable picture of the overall population structure across the three marker technologies (**Figure 2**). The total population clustered into a set of elite lines and PGRs. Similar findings have been reported by Cavanagh et al. (2013) investigating the diversity of 2,994 accessions of hexaploid wheat including landraces and modern cultivars and by Balfourier et al. (2019) examining the phylogeography of 4,506 landraces and cultivars originating from 105 different countries. Moreover, we observed that the estimates of the RD matrices using the array-based scoring of SNPs and GBS were similar, which is reflected by correlations for the total population of 0.83 (**Table 3**). This finding is in accordance with a previous study in wheat with U.S. elite lines (Elbasyoni et al., 2018) but also for other crops such as maize (e.g., Frascaroli et al., 2013) or barley genetic resources (Darrier et al., 2019). In contrast, the moderate correlations between distance matrices calculated based on SSR and GBS or SNP array markers (**Table 3**) are most likely caused by the limited number of SSR markers used in our study, which is in accordance with previous study in wild and cultivated barley (Hübner et al., 2012). This can be deduced from a high correlation ($r = 0.85$, $P < 0.01$) observed between kinship matrices calculated using a 90k SNP array and 782 SSR markers for 372 elite wheat lines observed in the study of Jiang et al. (2015). The low number of SSR markers, however, reflects comparable cost scenarios and shows that SSR markers are less suitable for large-scale characterization of wheat collections.

Use of Genome-Wide Prediction to Provide Detailed Information for Entire Wheat Collections

More than half a million wheat genetic resources are conserved worldwide in genebanks (Longin and Reif, 2014). Detailed information on their phenotypic diversity is lacking, but is necessary to enable a targeted selection of promising accessions for (pre-)breeding. In a proof-of-concept study in sorghum, Yu et al. (2016) demonstrated the potential to use genome-wide predictions to efficiently provide phenotypic information about entire genebank collections. Our study confirmed the results in wheat for the two important agronomic traits grain yield and HD (**Figure 3**). The high prediction ability can be explained by the large genetic variation in our study. The population we used contained about 50% of PGRs, with grain yields ranging from 4.75 to 10.14 Mg ha^{-1} (**Supplementary Figure 4**) and a genetic variance of 0.98 $(Mg\ ha^{-1})^2$. We observed four times higher genetic variance compared to elite wheat lines in Europe (He et al., 2017). Although the genetic structure of the traits influences the prediction accuracy, it is difficult to say if this was the main driving factor of the prediction ability in our study. The lower predictability for HD reported in our study is consistent with the study of Bentley et al. (2014). They used a similar population size with 376 European elite wheat lines (from France, Germany, and the UK) and reported the average prediction accuracy of flowering time (0.52) to be considerably lower than grain yield (0.68), despite the higher heritability of flowering time compared to yield. The choice of marker systems did not strongly influence the prediction abilities, except for the SSR markers, which is presumably mainly due to the low number of markers (Jiang et al., 2014). Our results are consistent with a recent study in wheat that contrasted the potential and limitations of array-based scoring of SNPs and GBS to perform genome-wide prediction (Elbasyoni et al., 2018). The combination of marker information with two- or three-kernel models slightly improved prediction ability (**Figure 3**) and represents a solid approach for populations genotyped with different marker platforms. Interestingly, we found that very low frequency markers contributed to the improvement of prediction ability (**Figure 4**). However, such markers are usually deleted as outliers in SNP arrays but can be reliably captured by GBS. The potential of rare alleles to improve prediction ability combined with the expectation that valuable novel diversity is most likely rare (Mascher et al., 2019) suggests that rare markers deserve careful consideration in the design of the pre-breeding program.

CONCLUSION

We observed an ascertainment bias for wheat caused by array-based SNP markers, which particularly impacts the estimates of the within population diversity. This was not the case with GBS, which makes it an interesting marker system to fingerprint entire genebank collections. In summary, our study showed the potential of genebank genomics to unlock the genetic diversity maintained in genebanks.

AUTHOR'S NOTE

All authors declare that this study adheres to standard biosecurity and institutional safety procedures.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

All authors declare that this study adheres to ethical standards including ethics committee approval and consent procedure. All experiments were performed under the current laws of Germany.

AUTHOR CONTRIBUTIONS

JC, YZ, and JR designed the study. NS and MR contributed to the generation of genomic data. JC, SB, AS, and NP curated phenotypic and genomic data. JC performed the analyses. JC and JR wrote the paper with input from all co-authors.

ACKNOWLEDGMENTS

The Federal Ministry of Education and Research of Germany is acknowledged for funding AS (grant no. FKZ031B0184B).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00042/full#supplementary-material

REFERENCES

Adu, G. B., Awuku, F. J., Amegbor, I. K., Haruna, A., Manigben, K. A., and Aboyadana, P. A. (2019). Genetic characterization and population structure of maize populations using SSR markers. *Ann. Agric. Sci.* 64, 47–54. doi: 10.1016/j.aoas.2019.05.006

Alipour, H., Bihamta, M. R., Mohammadi, V., Peyghambari, S. A., Bai, G., and Zhang, G. (2017). Genotyping-by-sequencing (GBS) revealed molecular genetic diversity of Iranian wheat landraces and cultivars. *Front. Plant Sci.* 8, 1293. doi: 10.3389/fpls.2017.01293

Allen, A. M., Barker, G. L. A., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., et al. (2011). Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnol. J.* 9, 1086–1099. doi: 10.1111/j.1467-7652.2011.00628.x

Balfourier, F., Bouchet, S., Robert, S., Oliveira, R. D., Rimbert, H., Kitt, J., et al. (2019). Worldwide phylogeography and history of wheat genetic diversity. *Sci. Adv.* 5, eaav0536. doi: 10.1126/sciadv.aav0536

Bentley, A. R., Scutari, M., Gosman, N., Faure, S., Bedford, F., Howell, P., et al. (2014). Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theor. Appl. Genet.* 127, 2619–2633. doi: 10.1007/s00122-014-2403-y

Bernal-Vasquez, A. M., Utz, H. F., and Piepho, H. P. (2016). Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. *Theor. Appl. Genet.* 129, 787–804. doi: 10.1007/s00122-016-2666-6

Bhatta, M., Morgounov, A., Belamkar, V., Poland, J., and Baenziger, P. S. (2018). Unlocking the novel genetic diversity and population structure of synthetic Hexaploid wheat. *BMC Genomics* 19, 591. doi: 10.1186/s12864-018-4969-2

Boeven, P. H. G., Zhao, Y., Thorwarth, P., Liu, F., Maurer, H. P., Gils, M., et al. (2019). Negative dominance and dominance-by-dominance epistatic effects reduce grain-yield heterosis in wide crosses in wheat. (in review).

Borlaug, N. E. (1968). "Wheat breeding and its impact on world food supply". *Third international wheat genetics symposium*. vol. 5-9. Ed. K. W. Finley and K. W. Sheppard (Canberra, ACT: Australian Academy of Science), 1–36.

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). ASREML-R Reference Manual. Release 3.0. Technical Report, Queensland Department of Primary Industries, Australia.

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

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15, 1496–1502. doi: 10.1101/gr.4107905

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., et al. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front. Plant Sci.* 10, 544. doi: 10.3389/fpls.2019.00544

Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci. J.* 270, 123–130. doi: 10.1016/j.plantsci.2018.02.019

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 (5), e19379. doi: 10.1371/journal.pone.0019379

Eltaher, S., Sallam, A., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F. M., et al. (2018). Genetic diversity and population structure of F3:6 Nebraska winter wheat genotypes using genotyping-by-sequencing. *Front. Plant Sci.* 9, 76. doi: 10.3389/fgene.2018.00076

FAO (2017). The future of food and agriculture - Trends and challenges. Food and Agriculture Organization of the United Nations. Rome.

Fischer, T., Byerlee, D., and Edmeades, G. (2014). Crop yields and global food security: will yield increase continue to feed the world? Australian Centre for International Agricultural Research, Canberra. http://aciar.gov.au/publication/mn158.

Frascaroli, E., Schrag, T. A., and Melchinger, A. E. (2013). Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. *Theor. Appl. Genet.* 126, 133–141. doi: 10.1007/s00122-012-1968-6

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325–338. doi: 10.1093/biomet/53.3-4.325

Hübner, S., Günther, T., Flavell, A., Fridman, E., Graner, A., Korol, A., et al. (2012). Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Mol. Ecol.* 21, 1115–1129. doi: 10.1111/j.1365-294X.2011.05434.x

He, S., Zhao, Y., Mette, M. F., Bothe, R., Ebmeyer, E., Sharbel, T., et al. (2015). Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L). *BMC Genomics* 16, 168. doi: 10.1186/s12864-015-1366-y

He, S., Reif, J. C., Korzun, V., Bothe, R., Ebmeyer, E., and Jiang, Y. (2017). Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. *Theor. Appl. Genet.* 130, 635–647. doi: 10.1007/s00122-016-2840-x

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* 15, 896–904. doi: 10.1038/s41588-019-0382-2

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L., and Sorrells, M. E. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PloS One* 8 (9), e74612. doi: 10.1371/journal.pone.0074612

Himmelbach, A., Knauft, M., and Stein, N. (2014). Plant sequence capture optimised for illumina sequencing. *Bio-protocol* 4 (13), e1166. doi: 10.21769/BioProtoc.1166

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *BRIEF Funct. Genomics* 9 (2), 166–177. doi: 10.1093/bfgp/elq001

Jiang, Y., Zhao, Y., Rodemann, B., Plieske, J., Kollers, S., Korzun, V., et al. (2015). Potential and limits to unravel the genetic architecture and predict the variation of Fusarium head blight resistance in European winter wheat (Triticum aestivum L.). *Heredity* 114 (3), 318–326. doi: 10.1038/hdy.2014.104

Jordan, K. W., Wang, S., Lun, Y., Gardiner, L.-J., MacLachlan, R., Hucl, P., et al. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on omoeologous genomes. *Genome Biol.* 16, 48. doi: 10.1186/s13059-015-0606-4

King, J., Grewal, S., Yang, C.-y., Hubbart, S., Scholefield, D., Ashling, S., et al. (2017). A step change in the transfer of interspecific variation into wheat from Amblyopyrum muticum. *Plant Biotechnol. J.* 15, 217–226. doi: 10.1111/pbi.12606

Kollers, S., Rodemann, B., Ling, J., Korzun, V., Ebmeyer, E., Argillier, O., et al. (2013). Whole genome association mapping of fusarium head blight resistance in European winter wheat (Triticum aestivum L.). *PloS One* 8 (2), e57500. doi: 10.1371/journal.pone.0057500

Laidig, F., Piepho, H.-P., Drobek, T., and Meyer, U. (2014). Genetic and non-genetic long-term trends of 12 different crops in German official variety performance trials and on-farm yield trends. *Theor. Appl. Genet.* 127, 2599–2617. doi: 10.1007/s00122-014-2402-z

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (6), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*. 1303.3997.

Liu, G., Zhao, Y., Gowda, M., Longin, C. F. H., Reif, J. C., and Mette, M. F. (2016). Predicting hybrid performances for quality traits through genomic-assisted approaches in central European wheat. *PloS One* 11 (7), e0158635. doi: 10.1371/journal.pone.0158635

Longin, C. F. H., and Reif, J. C. (2014). Redesigning the exploitation of wheat genetic resources. *Trends. Plant Sci.* 19 (10), 631–636. doi: 10.1016/j.tplants.2014.06.012

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27 (2), 209–220.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17.1, 10–12. doi: 10.14806/ej.17.1.200

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* 51, 1076–1081. doi: 10.1038/s41588-019-0443-6

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Har. Protoc.* 6, prot5448. doi: 10.1101/pdb.prot5448

Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., et al. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51, 319–326. doi: 10.1038/s41588-018-0266-x

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* 19, 318. doi: 10.1186/s12870-019-1926-4

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Philipp, N., Weichert, H., Bohra, U., Weschke, W., Schulthess, A. W., and Weber, H. (2018). Grain number and grain yield distribution along the spike remain stable despite breeding for high yield in winter wheat. *PloS One* 13 (10), e0205452. doi: 10.1371/journal.pone.0205452

Plaschke, J., Ganal, M. W., and Röder, M. S. (1995). Detection of genetic diversity in closely related bread wheat using microsatelUte markers. *Theor. Appl. Genet.* 91, 1001–1007. doi: 10.1007/BF00223912

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 (2), e32253. doi: 10.1371/journal.pone.0032253

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., and Bidet, Y. (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

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ .

Röder, M. S., Plaschke, J., König, S. U., Börner, A., Sorrells, M. E., Tanksley, S. D., et al. (1995). Abundance, variability and chromosomal location of microsatellites in wheat. *Mol. Gen. Genet.* 246, 327–333. doi: 10.1007/BF00288605

Röder, M. S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.-H., Leroy, P., et al. (1998). A microsatellite map of wheat. *Genetics* 149, 2007–2023.

Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F. C., He, Z., and Rajaram, S. (2018). Wheat genetic resources in the post-genomics era: promise and challenges. *Ann. Bot.* 121, 603–616. doi: 10.1093/aob/mcx148

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PloS One* 8 (6), e66428. doi: 10.1371/journal.pone.0066428

Riaz, A., Hathorn, A., Dinglasan, E., Ziems, L., Richard, C., Singh, D., et al. (2017). Into the vault of the Vavilov wheats: old diversity for new alleles. *Genet. Resour. Crop Evol.* 64, 531–544. doi: 10.1007/s10722-016-0380-5

Romay, M., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14, R55. doi: 10.1186/gb-2013-14-6-r55

Rufo, R., Alvaro, F., Royo, C., and Soriano, J. M. (2019). From landraces to improved cultivars: assessment of genetic diversity and population structure of Mediterranean wheat using SNP markers. *PloS One* 14 (7), e0219867. doi: 10.1371/journal.pone.0219867

Sajjad, M., Khan, S. H., and Shahzad, M. (2018). Patterns of allelic diversity in spring wheat populations by SSR-markers. *Cytol. Genet.* 52 (2), 155–160. doi: 10.3103/S0095452718020081

Schulthess, A. W., Reif, J. C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., et al. (2017). The roles of pleiotropy and close linkage as revealed by association mapping of yield and correlated traits of wheat (Triticum aestivum L.). *J. Exp. Bot.* 68 (15), 4089–4101. doi: 10.1093/jxb/erx214

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 (7), e0132112. doi: 10.1371/journal.pone.0132112

Singh, N., Wu, S., Raupp, W. J., Sehgal, S., Arora, S., Tiwari, V., et al. (2019). Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep.* 9, 650. doi: 10.1038/s41598-018-37269-0

The International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 345 (6194), 1251788. doi: 10.1126/science.1251788

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vikram, P., Franco, J., Burgueño-Ferreira, J., Li, H., Sehga, D., Pierre, C. S., et al. (2016). Unlocking the genetic diversity of Creole wheats. *Sci. Rep.* 6, 23092. doi: 10.1038/srep23092

Voss-Fels, K. P., Stahl, A., Wittkop, B., Lichthardt, C., Nagler, S., Rose, T., et al. (2019). Breeding improves wheat productivity under contrasting agrochemical input levels. *Nat. Plants* 5, 706–714. doi: 10.1038/s41477-019-0445-5

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 Biotech. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9

Wang, Y., Rashid, M. A. R., Li, X., Yao, C., Lu, L., Bai, J., et al. (2019). Collection and evaluation of genetic diversity and population structure of potato landraces and varieties in China. *Front. Plant Sci.* 10, 139. doi: 10.3389/fpls.2019.00139

Wendler, N., Mascher, M., Nöh, C., Himmelbach, A., Scholz, U., Ruge-Wehling, B., et al. (2014). Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol. J.* 12, 1122–1131. doi: 10.1111/pbi.12219

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., et al. (2004). Diversity arrays technology (DArT) for whole-genome profiling of barley. *Proc. Natl. Acad. Sci. U.S.A.* 101 (26), 9915–9920. doi: 10.1073/pnas.0401076101

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L. A., 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

Wingen, L. U., Orford, S., Goram, R., Leverington−Waite, M., Bilham, L., Patsiou, T. S., et al. (2014). Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat. *Theor. Appl. Genet.* 127, 1831–1842. doi: 10.1007/s00122-014-2344-5

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., et al. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants* 2, 16150. doi: 10.1038/nplants.2016.150

Zanke, C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al. (2014a). Genetic architecture of main effect QTL for heading date in European winter wheat. *Front. Plant Sci.* 5, 1–12. doi: 10.3389/fpls.2014.00217

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al. (2014b). Whole genome association mapping of plant height in winter wheat (Triticum aestivum L.). *PloS One* 91 (11), e113287. doi: 10.1371/journal.pone.0113287

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al. (2015). Analysis of main effect QTL for thousand grain weight in European winter wheat (TriticumaestivumL.) by genome-wide association mapping. *Front. Plant Sci.* 6, 644. doi: 10.3389/fpls.2015.00644

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 Plant Science

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/fpls.2020.00176

Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants

Claudius Grehl[1,2*], *Marc Wagner*[3], *Ioana Lemnian*[1,4], *Bruno Glaser*[2] *and Ivo Grosse*[1,5]

[1] *Institute of Computer Science, Bioinformatics, Martin Luther University Halle–Wittenberg, Von Seckendorff-Platz 1, Halle (Saale), Germany,* [2] *Institute of Agronomy and Nutritional Sciences, Soil Biogeochemistry, Martin Luther University Halle–Wittenberg, Von Seckendorff-Platz 3, Halle (Saale), Germany,* [3] *Institute of Mathematics and Informatics, Freie Universität Berlin, Berlin, Germany,* [4] *Institute of Human Genetics, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany,* [5] *Bioinformatics Unit, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany*

OPEN ACCESS

Edited by:
Sebastian Beier,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
R. Keith Slotkin,
Donald Danforth Plant Science Center, United States
Kaushik Panda,
Danforth Plant Science Center, United States
in collaboration with reviewer RKS,
David Roquis,
Technical University of Munich, Germany

****Correspondence:***
Claudius Grehl
claudius.grehl@informatik.uni-halle.de

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *12 October 2019*
Accepted: *05 February 2020*
Published: *28 February 2020*

Citation:
Grehl C, Wagner M, Lemnian I, Glaser B and Grosse I (2020) Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. Front. Plant Sci. 11:176. doi: 10.3389/fpls.2020.00176

DNA methylation is involved in many different biological processes in the development and well-being of crop plants such as transposon activation, heterosis, environment-dependent transcriptome plasticity, aging, and many diseases. Whole-genome bisulfite sequencing is an excellent technology for detecting and quantifying DNA methylation patterns in a wide variety of species, but optimized data analysis pipelines exist only for a small number of species and are missing for many important crop plants. This is especially important as most existing benchmark studies have been performed on mammals with hardly any repetitive elements and without CHG and CHH methylation. Pipelines for the analysis of whole-genome bisulfite sequencing data usually consists of four steps: read trimming, read mapping, quantification of methylation levels, and prediction of differentially methylated regions (DMRs). Here we focus on read mapping, which is challenging because un-methylated cytosines are transformed to uracil during bisulfite treatment and to thymine during the subsequent polymerase chain reaction, and read mappers must be capable of dealing with this cytosine/thymine polymorphism. Several read mappers have been developed over the last years, with different strengths and weaknesses, but their performances have not been critically evaluated. Here, we compare eight read mappers: Bismark, BismarkBwt2, BSMAP, BS-Seeker2, Bwameth, GEM3, Segemehl, and GSNAP to assess the impact of the read-mapping results on the prediction of DMRs. We used simulated data generated from the genomes of *Arabidopsis thaliana*, *Brassica napus*, *Glycine max*, *Solanum tuberosum*, and *Zea mays*, monitored the effects of the bisulfite conversion rate, the sequencing error rate, the maximum number of allowed mismatches, as well as the genome structure and size, and calculated precision, number of uniquely mapped reads, distribution of the mapped reads, run time, and memory consumption as features for benchmarking the eight read mappers mentioned above. Furthermore, we validated our findings using real-world data of *Glycine max* and showed the influence of the mapping step on DMR calling in WGBS pipelines. We found that the conversion rate had only a minor impact on the mapping quality and the number of uniquely mapped reads, whereas the error rate and the maximum number of

allowed mismatches had a strong impact and leads to differences of the performance of the eight read mappers. In conclusion, we recommend BSMAP which needs the shortest run time and yields the highest precision, and Bismark which requires the smallest amount of memory and yields precision and high numbers of uniquely mapped reads.

Keywords: epigenetics, DNA methylation patterns, read mapping, benchmarking, WGBS

INTRODUCTION

It has been shown that DNA methylation is involved in several biological mechanisms and diseases such as cancer (Koch et al., 2018). Plant methylation analysis is especially interesting as 5-methyl-cytosine (5mC) is involved in the heterosis effect (Chen et al., 2018), transposon silencing, and environment-dependent transcriptome plasticity (Lauss et al., 2018). However, in addition to the complementary CG methylation being highly abundant in animals, in plants CHG and uncomplimentary CHH (H=C,T or A) methylation have evolved from the former recognition system of foreign DNA.

Whole-genome bisulfite sequencing (WGBS) is often referred to as the "gold standard" for 5mC detection because the whole genome is covered at a single-base resolution. Other methods cover only preselected genome regions enriched for cytosine-phosphate-guanine-dinucleotide (CpG) content or methylation, for example with the use of restriction enzymes in reduced representation bisulfite sequencing (rrBS) (Sun et al., 2015), or methylated DNA immune precipitation, followed by next generation sequencing (MeDIP-seq) (Bock et al., 2010; Aberg et al., 2017).

Bisulfite-mediated conversion of unmethylated cytosines to uracil, and during PCR to thymine, leads to four different strands in the data sets after sequencing: original top, complementary to original top, original bottom, and complementary to original bottom strand (**Figure 1**). Methylated cytosines remain unaffected and could be spotted by alignment of the generated sequencing reads to a reference genome or a non-bisulfite-treated control.

Critical within the bioinformatics analysis of WGBS data sets is the mapping step, as the reduced alphabet leads to specific challenges for the mapping tools due to the bisulfite treatment (Laird, 2010).

In general, two different algorithmic approaches exist in bisulfite-read alignment tools for dealing with the unmethylated C to T conversion: the 'wild card' and the 'three letter' approach. In the wild card approach the Cs in the reference genome are replaced with the wild card 'Y' for pyrimidine bases and thus allows for the alignment of Cs (methylated Cs) and Ts (possibly unmethylated Cs). The alignment itself is based on matching seeds (k-mers) to the reference and then extending them. In the three letter approach the alphabet of the genome and the reads is reduced to {A, G, T}, by converting all Cs in the reference sequence and in the read data to Ts. Afterwards, the reads are mapped by conventional mappers such as Bowtie, Bowtie2, or bwa, so the alignment of bisulfite data profits directly from the improvements of traditional mappers.

One study already focused on the benchmarking of rrBS alignment using simulated rrBS and real human lung tissue data sets (Sun et al., 2018). Kunde-Ramamoorthy et al. (2014) evaluated the mapping performance of five mapping tools in WGBS datasets of human peripheral blood lymphocyte and a hair follicle. Another study has been performed in plants, showing that the tool BismarkBwt2 performed best in terms of sensitivity and precision, not accounting for the coverage distribution across the reference genome (Omony et al., 2019). In contrast, our study is focused on simulated WGBS in plants, covering different species with a different amount of repetitive sequences. In addition, little is known about the mapping behavior in crop plants. Furthermore, as all former studies did not systematically account for different parameter settings such as the number of mismatches, we evaluated this parameter in more detail.

There is need for an extensive qualitative and quantitative benchmarking of alignment tools, to avoid the false interpretation of results in DNA methylation studies and to enable the application of precise, efficient, and user-friendly

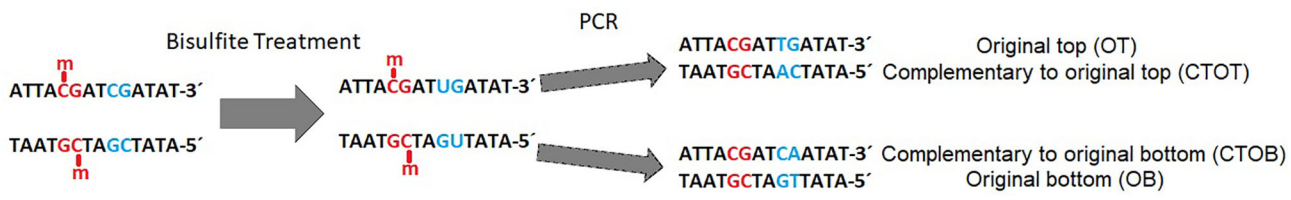

FIGURE 1 | Example of a DNA double strand with methylated (red) and unmethylated (blue) CpG-site (cytosine-phosphate-guanine-dinucleotide) before and after bisulfite treatment and polymerase chain reaction (PCR). Methylated cytosine is not affected by bisulfite, whereas unmethylated cytosine is converted to uracil and further on to thymine during PCR in the original top strand, and to adenine in the complementary top strand [adapted from (Grehl et al., 2018)].

pipelines. The known "truth" is especially important, and this could be generated by benchmarking datasets of simulated and, thus, known read data, to calculate the quality of scores in multiclass hypothesis testing. In terms of quantitative comparison, time efficiency, amount of uniquely mapped reads, and consumption of RAM has to be monitored, as well as the overall distribution of mapped reads, to look for genomic regions with systematically lower coverages.

MATERIAL AND METHODS

Arabidopsis thaliana, *Brassica napus*, *Glycine max*, *Solanum tuberosum*, and *Zea mays* (**Table 1**) have been examined to reveal potential inter-species variability in terms of mappability. These species have been chosen to cover different agronomically relevant plant families, different genome sizes, and different assembly qualities. All reference genomes were downloaded from http://plants.ensembl.org. We simulated WGBS datasets for 2 x 150 base pair paired-end reads for the five different plant genomes, using the open-source WGBS simulation tool: Sherman (https://www.bioinformatics.babraham.ac.uk/projects/sherman/), which has been developed at the Babraham Institute. The reads have been simulated. We chose 150bp paired-end sequencing for our benchmarking study as it is the mostly applied and proposed sequencing option for WGBS experiments. In doing so, small repeats below the total fragment size of 500-700bp could be covered, which is especially important for repeat-rich (crop) plants. Furthermore, choosing a parameter set of 150bp paired end, facilitates the necessary multiplexing with non-bisulfite libraries during sequencing.

For each species, benchmarking datasets in 5-fold sequencing depth, three bisulfite conversion rates [90%, 98%, 100%], and four different sequencing error rates [0%, 0.1%, 0.5%, 1%] were simulated. The sequencing errors were modelled to account for more likely errors at the end of a read, like in real world sequencing data (**Figure 2**). Whereas the overall resulting phred score of 30 is equivalent to an error rate of 0.1% or 1 in 1000 wrong base calls. Illumina HiSeq sequencing yields an error rate of 0.0034-1% while PacBio shows 5–10% false base calls (Escalona et al., 2016). We decided to include a 98% conversion rate as this is usually guaranteed by sequencing facilities, and 90% to look for a value below this threshold.

For mapping the simulated WGBS reads to the genomes, we tested several wild-card and three-letter mappers: Bismark (Krueger and Andrews, 2011), BSMAP (Xi and Li, 2009), BS-Seeker2 (Guo et al., 2013), Bwa-meth (Pedersen et al., 2014), GEM3 (Marco-Sola et al., 2012), GSNAP (Wu and Nacu, 2010), and Segemehl (Hoffmann et al., 2009; Otto et al., 2012). These mappers differ in terms of their "age", number of citations, and indexing strategy (**Table 2**). For further insight into mapping and indexing strategies, as well as for an insight into the underlying algorithmic approaches we recommend (Tran et al., 2014).

Bismark (Krueger and Andrews, 2011), one of the most cited three letter mapper for bisulfite-sequencing data, first converts the reads and the genome into two versions: a C-to-T and a G-to-A version. Afterwards, the two read versions are aligned to the two versions of the reference genome with the goal of detecting to which of the four strands (**Figure 1**) the read fits. This alignment is performed by four parallel instances of either Bowtie (Langmead et al., 2009), one of the fastest mappers for NGS data, or Bowtie2 (Langmead and Salzberg, 2012), an improved version of bowtie, that allows gapped alignment.

BSMAP (Xi and Li, 2009) is included in the list for being the first mapper for the alignment of bisulfite data. It uses an efficient HASH table the seeding algorithm for indexing the genome, bitwise masking each nucleotide in the reads and the reference and matching them to each other in an efficient way.

GSNAP (Wu and Nacu, 2010) is a general purpose mapper that can also deal with bisulfite data. Like BSMAP, it is based on special hash tables and uses a wild-card approach to match read seeds to genome regions. Since its original publication several improvements of the algorithms have been made by increasing the length of the hashed k-mers, adding a compression mechanism and using enhanced suffix arrays (Wu et al., 2016).

BS-Seeker2 (Guo et al., 2013) is the extension of BS Seeker (Chen et al., 2010) for mapping bisulfite data and deploys a three letter approach. In addition to performing a gapped alignment it can filter out reads with incomplete bisulfite conversion, in this way increasing the specificity.

Compared to the other tools in this benchmark Bwa-meth (Pedersen et al., 2014) is a relatively new mapper for bisulfite data. It is based on BWA-mem aligner (Li and Durbin, 2009; Li and Durbin, 2010) and it is advertised as a fast and accurate aligner.

Segemehl was originally designed as a general purpose mapper (Hoffmann et al., 2009) but has been extended to handle bisulfite data (Otto et al., 2012). Segemehl achieves a high sensitivity by using a wild-card approach based on enhanced suffix arrays for the seed search and the Myers bit-vector algorithm for computing semi-global alignments.

TABLE 1 | Five species included in this benchmarking study with the size of the reference genome and the used reference genome version which has been taken for the simulation of the read datasets.

Species	*Arabidopsis thaliana*	*Brassica napus*	*Glycine max*	*Solanum tuberosum*	*Zea mays*
Genome Size (bp)	135,670,229	738,357,821	955,377,461	727,424,546	2,104,350,183
Genome version	TAIR10	AST_PRJEB5043_v1	Glycine_max_v2.1	SolTub_3.0	B73 RefGen_v4
Repeats in %	<23	~48	~57	~49	~75
	(Flutre et al., 2011)	(Liu et al., 2018)	(Schmutz et al., 2010)	(Mehra et al., 2015)	(Wolf et al., 2015)
Citation	Lamesch et al., 2012	Chalhoub et al., 2014	Schmutz et al., 2010	Xu et al., 2011	Schnable et al., 2009

The proportion of repetitive sequences is given as the estimated value over the genome.

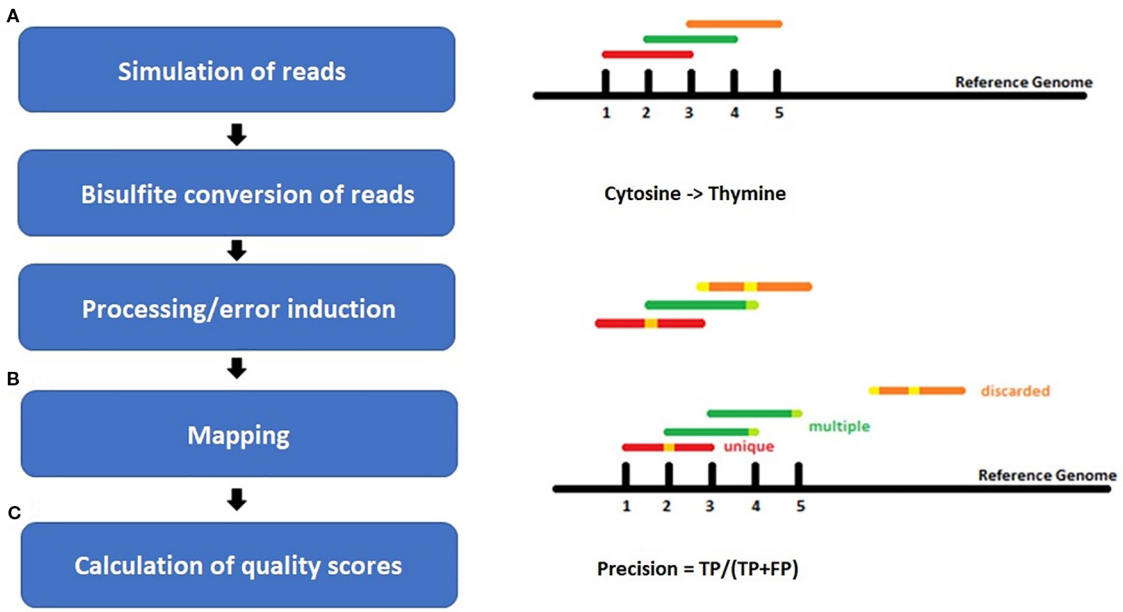

FIGURE 2 | Workflow of the experiment setup for **(A)** simulation of bisulfite-treated reads based on a reference genome using the tool Sherman, including bisulfite conversion and error induction. Afterwards **(B)** mapping of the simulated datasets and **(C)** calculation of quality scores. The color coding shows the different classes of reads after mapping: red = uniquely mapped, green multiply mapped, orange= discarded/unmapped reads.

TABLE 2 | Bisulfite Read mapping tools evaluated in this survey, listed by their mapping and indexing strategy.

Mapper name	Strategy	Indexing	Version	Citations
Bismark	3 letter	BWT (bowtie 2/bowtie 1)	0.19.1	1.176
BSMAP	wild-card	Hash table (SOAP)	2.73	3
BS-Seeker2	3 letter	BWT (bowtie 2)	2.1.5	135
Bwa-meth	3 letter	BWA mem	0.2.2	3
GEM3	3 letter	Custom FM index	3.6.1-2	236
GSNAP	wild-card	Hash table	2019-06-10	83
Segemehl	wild-card	Enhanced suffix array	0.2.0	283

The number of citations is based on the Web of Science Core Collection (date: 19.1.2020).

The eight mappers were used to map the simulated reads with 0, 1, 2, and 3 mismatches allowed in the read. As Bowtie2 and Bwa-meth do not allow setting the total number of mismatches in a read as a parameter, but in the seed instead, we performed our analysis on the basis of 0 mismatches in the seed for these mapping tools. Other parameter settings, such as the number of threads used, were the default values of the mentioned tools if not stated otherwise and are comparable across the different tools. All scripts are available at git-hub (https://github.com/grehl/benchWGBSmap).

After mapping, the reads can be classified into three different classes: i) discarded reads that could not be mapped, ii) multiple mapped reads that could be aligned but to more than one position on the reference genome because of sequence similarities, and iii) uniquely mapped reads, which have been mapped to one position only.

For further evaluation, we used only uniquely mapped reads. Since we did not account for insertions and deletions, we have considered only the first base of the read at its genome position. When calculating the quality scores, we have compared the true and the predicted position of a read. For each read the true genome origin is known, since Sherman encodes it in the read name, while the predicted position is derived from the alignment files.

The quality scores computed are the amount of unique reads considered, the precision, the memory consumption, and the time consumption of the tools. Furthermore, we looked at the read distribution over the reference genome to account for systematic mapping deficiencies.

The precision of a mapping tool for a data set has been computed using the formula for macro-averaged precision (macroAvgPrecision)(TP = true positives, FP = false positives):

$$macroAvgPrecision = \frac{\sum_{i=1}^{N}\left(\frac{TP_i}{TP_i+FP_i}\right)}{N}$$

We first calculated the precision for every position i, summed over all positions and divided by the total number of positions N. The macro-averaging was chosen as it weights FP higher than in the micro-averaging calculation of the precision. We furthermore used "precision" in this manuscript instead of "macro-averaged precision".

To evaluate the impact of the tested tools on DMR calling and to show the reliability of our simulated benchmark study, we included a real-world dataset of *Glycine max* root hair samples grown under 25°C and 40°C (Hossain et al., 2017).

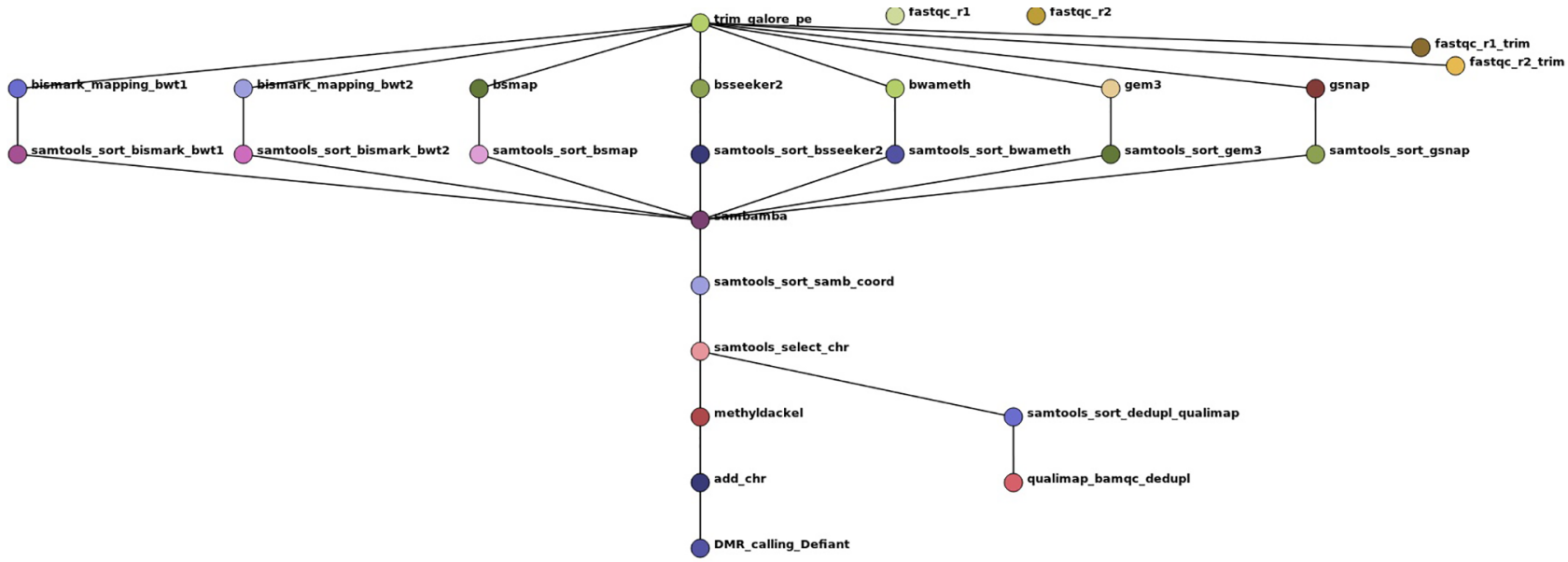

FIGURE 3 | Workflow show as Directed Acyclic Graph (DAG) of the benchmark experiments using 7 mapping tools on a real-world dataset of *Glycine max* root hair grown under two different temperatures to exemplarily compare the mapping performance and the influence of the mapping tools on DMR calling. The workflow includes two quality checks on the raw reads prior and after trimming with trim galore using fastqc, the mapping, a samtools sorting step to account for different output formats of the tools, a sambamba deduplication step, coordinate sorting, exclusion of scaffold mappings, methylation level calling using methyldackel and after adding "chr" to the methylation report output the final DMR calling with Defiant. To evaluate the bam quality we used the qualimap bamqc function.

For automatization we implemented a snakemake pipeline (Köster and Rahmann, 2012), shown in **Figure 3**. Other tools used in this pipeline are: trim galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), fastqc (Leggett et al., 2013), qualimap (García-Alcalde et al., 2012), samtools (Li et al., 2009), sambamba (Tarasov et al., 2015), methyldackel (https://github.com/dpryan79/MethylDackel) Defiant (v.1.1.6) (Condon et al., 2018) [parameter settings: -c 10 -v 'BY' -CpN 5 -p 0.05 -P 10], and Circos (Krzywinski et al., 2009). The mapping parameter sets were comparable and allowed 0 mismatches. As Segemehl showed an extensive memory consumption and runtime we had to exclude this mapper for the qualitative benchmark study and the mapping of the real dataset. For the DMR calling we had to exclude BS-Seeker2, as the flag information did not follow standard formats, so the files could not reliably be used for methylation calling. The settings for DMR calling were: minimum 10-fold coverage, minimum 5 CpN in one DMR, and minimum +/-10% methylation difference between the two treatments with a maximum p-value of 0.05. Analogous to the source code of the simulated benchmark study, we mainly relied on the default parameters if not stated otherwise in the script. All scripts are available online at git-hub (https://github.com/grehl/benchWGBSmap).

Simulation, mapping, and quality score calculation was performed on the IANVS High-Performance-Cluster of Martin-Luther University Halle-Wittenberg (**Table 3**). For calculation of runtime and memory consumption only one core of the login-nodes was allowed for mapping. For simplicity, to give an overview about mapping, and as the genome size is the most important factor with respect to runtime and memory consumption, we decided to focus on this factor only. A subsequent study could focus on the influence of other factors of runtime and memory consumption. The mapping of the quality benchmark was performed on the "large" and "small" nodes.

RESULTS

The results of our quantitative benchmark studies for memory consumption (**Figure 4**) and runtime (**Figure 5**) are shown for the eight mappers in relation to the size of the reference genome. The memory consumption ranged from 0.1 GB for the mapping of the *Arabidopsis thaliana* dataset with Bismark, either using bowtie or bowtie2, to 39 GB for the mapping of the *Zea mays* dataset with Segemehl. All datasets had a 100% conversion rate, a 0% error rate and, 0 mismatches were allowed during the mapping. Similar patterns in the memory consumption and runtime have also been observed for datasets with other parameter settings. In terms of runtime, the user time is depicted, ranging from a few minutes for all mappers using the *Arabidopsis thaliana* dataset, to 79 h for the mapping of a *Zea mays* dataset with Segemehl. Overall BSMAP took the least time, especially for large reference genomes. It is interesting to note that although *Solanum tuberosum* and *Brassica napus* have a similar genome size, and some mappers had a higher memory consumption (Segemehl, GEM3, GSNAP) and runtime (Segemehl, BismarkBwt2, GSNAP) for *Brassica napus*. This might be due to the large amount of large repetitive regions and paralogue genes within the *Brassica napus* genome, as the overall proportion of repeats is comparable to *Solanum tuberosum*.

Because of the extensive memory consumption and runtime of Segemehl, we excluded this mapper from the quality benchmark study.

Overall, the conversion rate did not influence the number of uniquely mapped reads or the mapping quality (**Supplementary Material**). In terms of the mapping quality, in relation to the error rate, and the reference genome, we basically see three groups of mappers (**Figure 6**). The first group is independent of the allowed number of mismatches during the mapping and includes Bismark, BismarkBwt2, Bwa-meth, and GEM3. The second group consists of BSMAP and BS-Seeker2, showing an increase in the number of uniquely mapped reads with higher numbers of allowed mismatches with barely any changes in precision. The third group, including GSNAP, shows an increase in the number of uniquely mapped reads but a decrease in the precision, with a higher number of mismatches allowed. As BismarkBwt2 and bwameth do not allow setting the number of mismatches in the entire read, both are labelled with a triangle. Between the analyzed genomes we see differences for all mappers with the tendency to lower numbers of uniquely mapped reads in *Zea mays* and lower precision in *Zea mays* and *Brassica napus*.

For *Arabidopsis thaliana* (**Figure 7**) and *Glycine max* (**Figure 8**) the distribution of reads over the reference genome is exemplarily

TABLE 3 | IANVS Cluster Specifications.

Node type	SLURM partition*	Qty.	CPU	Cores (total)	SMT threads (total)	Clock speed (GHz)	RAM (GiB)	Storage	InfiniBand blocking factor**	Remarks
login	–	2	2x12-core Intel Xeon	24	48	2.50	256	GPFS	1:8	–
small	standard	180	E5-2680v3				128	over IB	1:2 or 1:8	hostnames: small[001-180]
large		76					256			hostnames: large[001-076]
gpu	gpu	12								hostnames: gpu[01-12]
special	Special	2	4x 10-core E5-4620v3	40	80	2.00	1024		1:8	hostnames: special001, special 002

a "partition" in SLURM terms means "a group of machines to which one may submit cluster jobs".

***Nodes on the same InfiniBand switch always have their full bandwidth available when communicating with each other. However, if nodes want to communicate over switch boundaries, their available bandwidth might be reduced due to contention on the switch. The "blocking factor" is the maximum reduction of bandwidth that can occur in a case like this.*

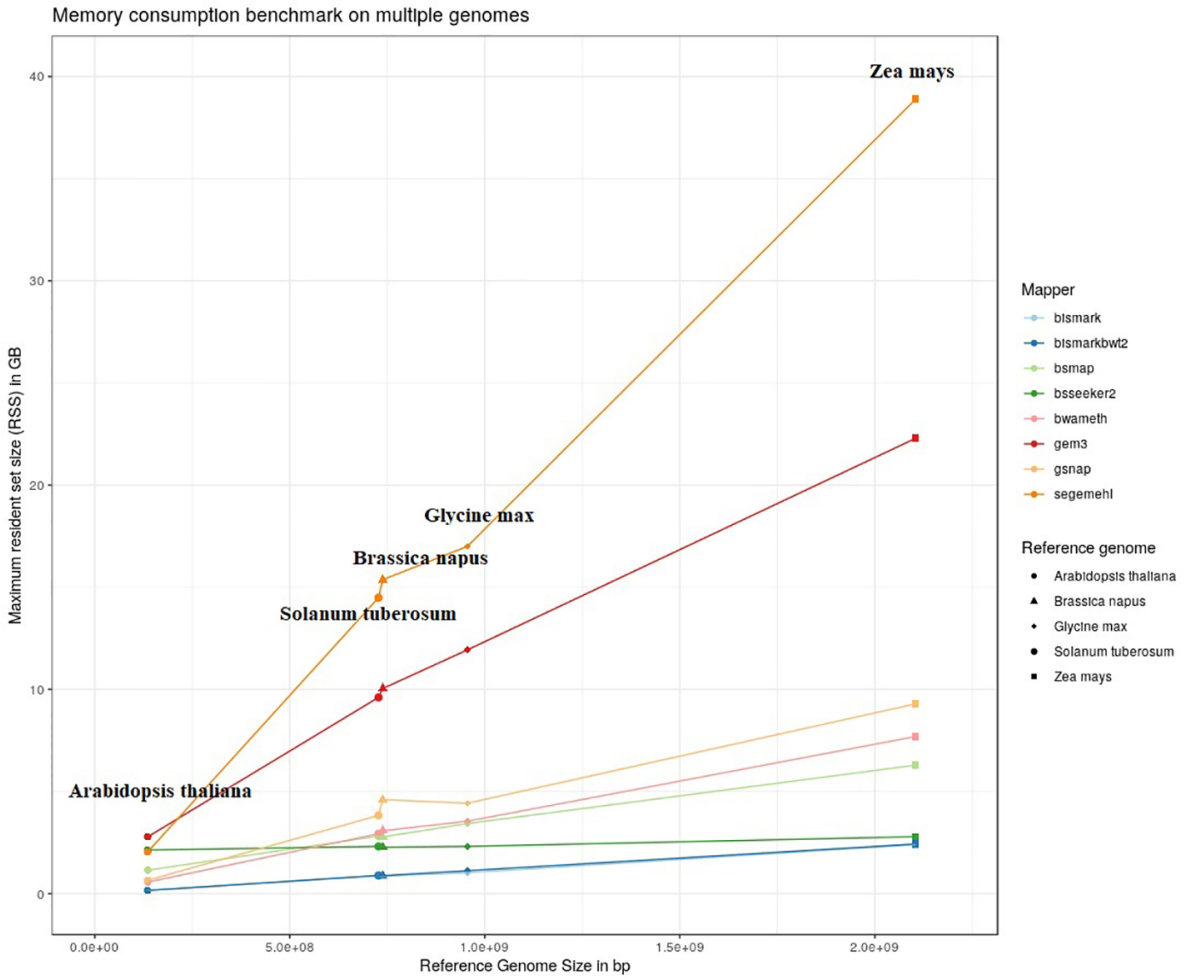

FIGURE 4 | Maximum resident set size in GB of 8 mappers for 5fold simulated bisulfite converted datasets out of five reference genomes (*Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum, Zea mays*). 0 % error rate, 100% conversion rate, 0 mismatch allowed.

shown for the mapping of two datasets each. The first dataset was simulated with a 100% conversion rate, a 0% error rate and was mapped with 0 mismatches allowed for all seven mapping tools, depicted in the lower window. The upper window shows a 100% conversion rate, a 1% error rate and 0 mismatches, again for all seven mapping tools. All coverage plots have a resolution of 400 windows across the whole reference genome. For higher error rates BSMAP, BS-Seeker2, and GSNAP show a severe decrease in coverage. Furthermore, we clearly see several regions with a decrease in coverage within the reference genome, independent of the error rate. In grey, we highlighted the regions which are known to contain a high percentage of repetitive sequences. Bismark and BismarkBwt2 are depicted behind each other, showing nearly the same coverage distribution. In total, Bwa-meth shows the least derivation in the coverage distribution.

The benchmarking of the real *Glycine max* dataset resulted in proper mapped paired end read counts (**Table 4**). The last column shows the final number of DMRs. These are additionally depicted in **Figure 9**.

DISCUSSION

We performed an extensive benchmarking experiment based on simulated data to evaluate the qualitative and quantitative performance of mappers for bisulfite sequencing data in five plant species with a focus on crop plants.

In terms of user time and memory consumption, the different tools showed big differences. Especially for larger genomes. For example Segemehl used a tremendous amount of RAM and needed the most time to map the given reads onto the reference genome. For larger reference genomes (>4 GB) the genome has to be split if Segemehl needs to be used. For these two reasons, we could not use

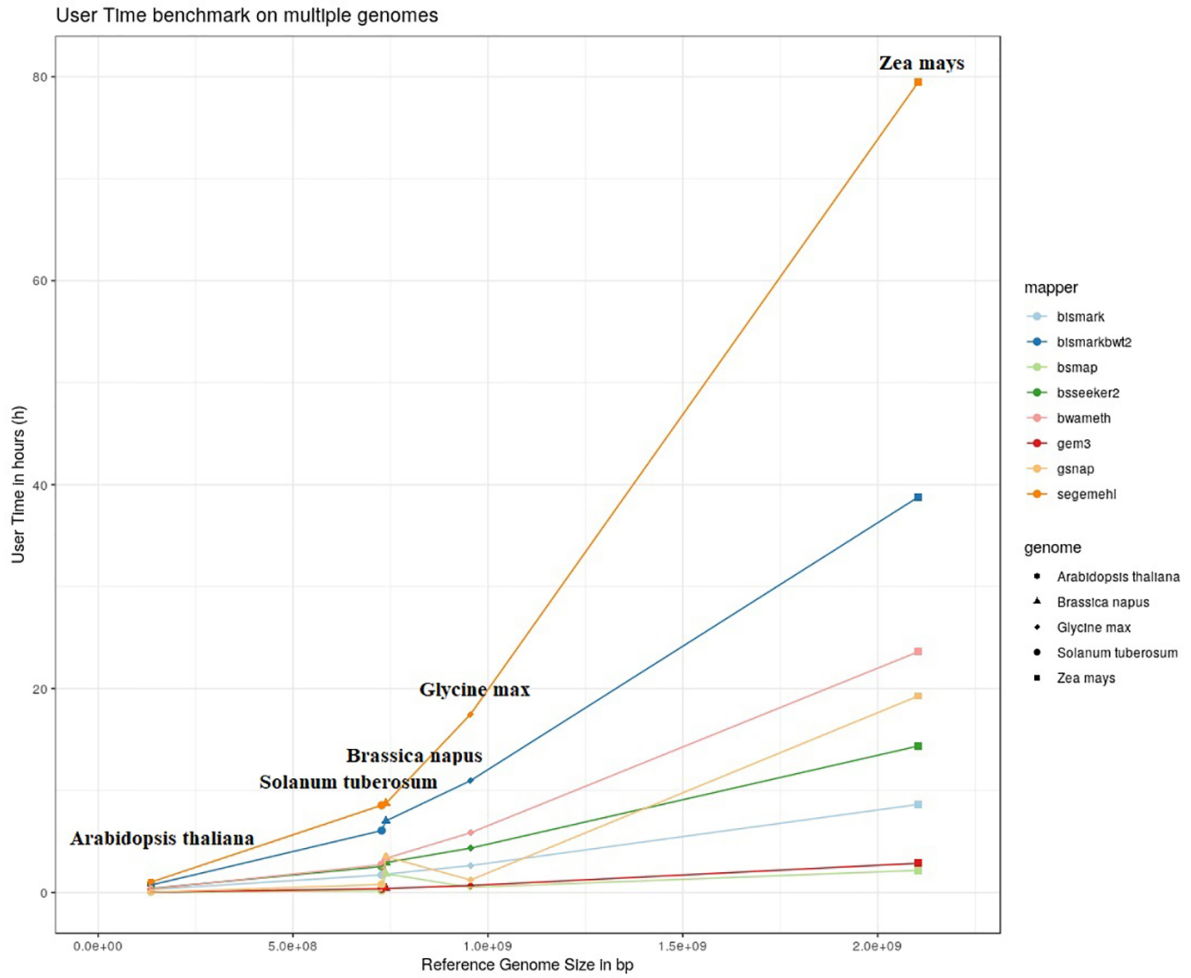

FIGURE 5 | User timer in hours of 8 mappers for 5fold simulated bisulfite converted datasets out of five reference genomes (*Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum, Zea mays*). 0 % error rate, 100% conversion rate, 0 mismatch allowed.

Segemehl for mapping of huge datasets such as *Zea mays*, even if it performed well in terms of precision in a pilot study. BSMAP, GEM3, and GSNAP showed only a low increase in time with the increasing size of the genome but used more memory. Bismark in particular showed a low increase for the memory consumption and a relatively low increase in run time. The large difference between Bismark and BismarkBwt2 is most likely due to the "soft clipping" function of BismarkBwt2.

The mapping quality and number of uniquely mapped reads change between the tools with *Zea mays* showing the lowest precision scores and the least number of uniquely mapped reads. This effect might be caused by the high number of repetitive sequences, which has been shown to make up to 75% of the *Zea mays* genome, containing mostly gypsy- and copia-like long, terminal repeats (LTR) (Wolf et al., 2015). For *Glycine max* the described number of repeats lays around ~57%. This also includes telomeric as well as centromeric repeats and not annotated repeats where the reference genome shows scaffolded regions (Schmutz et al., 2010). A wild-type reference genome sequencing consortium recently found 54% repeats (Xie et al., 2019). As most repeats are <50 bp (Sherman-Broyles et al., 2014), the 2 x 150 bp paired-end reads with an insert size of 200 bp – 400 bp could cover large parts of the genome uniquely. The distribution of reads across the reference genome shows a good overlap with known and long, repeat-rich regions. Some mappers such as GEM3 and GSNAP tend to map high amounts of FP in these regions. Other mappers leave these regions out, leading to a lower coverage.

In terms of precision, runtime and power to detect CpG-sites, Sun et al. (2018) found Bwa-meth and BS-Seeker2 to be the best tools based on simulated and real rrBS reads from human lung tumor tissue. However, this stands in contrast to our findings, which show precision deficiencies for Bwa-meth, with error rates above 0.1%, especially in repeat-rich and large plant genomes. BS-Seeker2

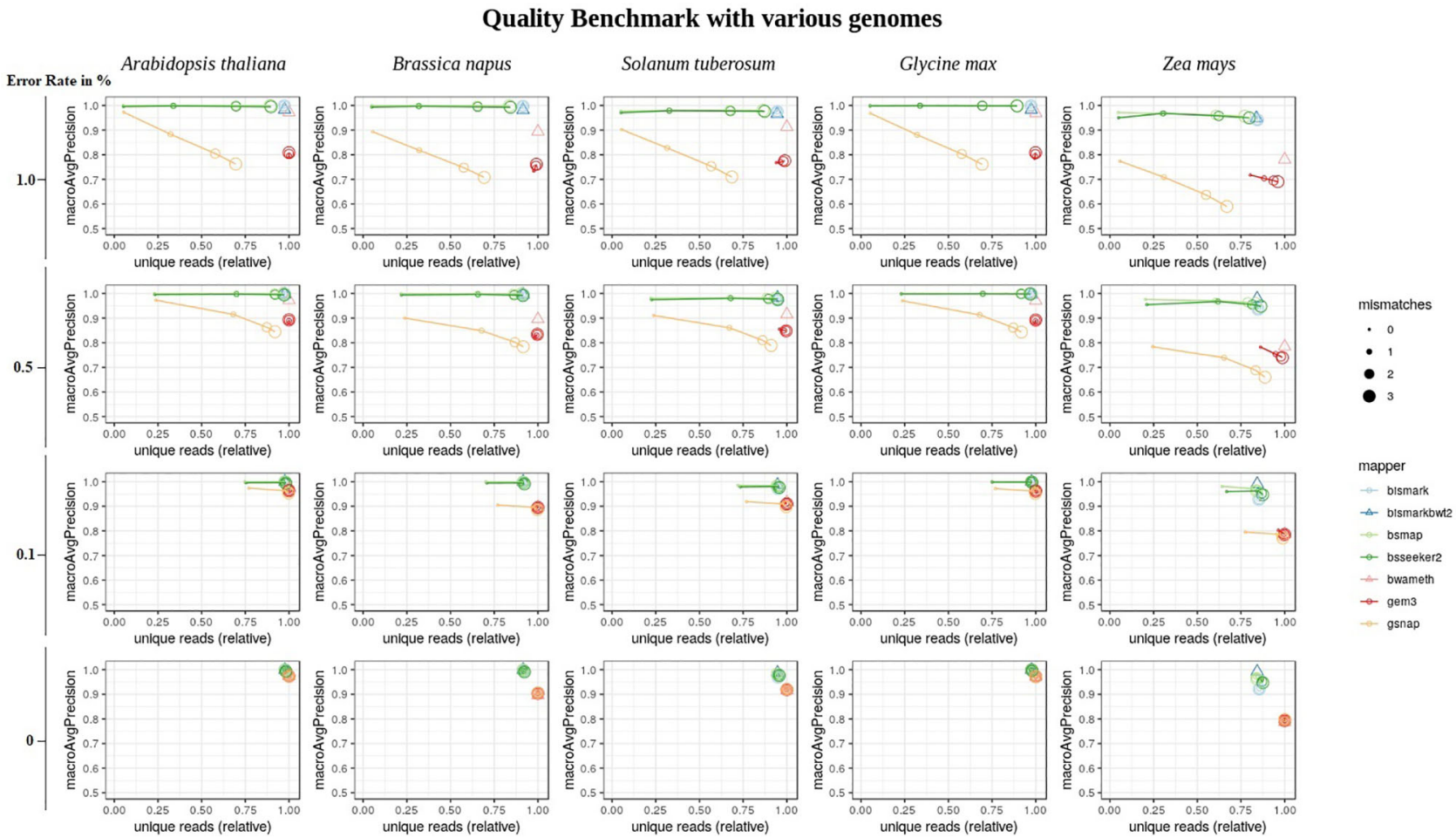

FIGURE 6 | Quality benchmark of 7 mappers based on simulated bisulfite sequencing datasets in *Arabidopsis thaliana, Brassica napus, Glycine max, Solanum tuberosum,* and *Zea mays.* We simulated the datasets with 4 different error rates [0, 0.1, 0.5 and 1 %] in a 5fold coverage. For 5 out of 7 mappers we had the opportunity to allow for different numbers of mismatches [0, 1, 2, 3]. These mappers are depicted by circles. Two mappers, bismark using bowtie2 and bwameth, did not allow the adjustment for different numbers of mismatches in the entire read. They are depicted by triangles. The conversion rate had no effect and is therefore not shown in this figure. The depicted conversion rate is 100% for all data sets.

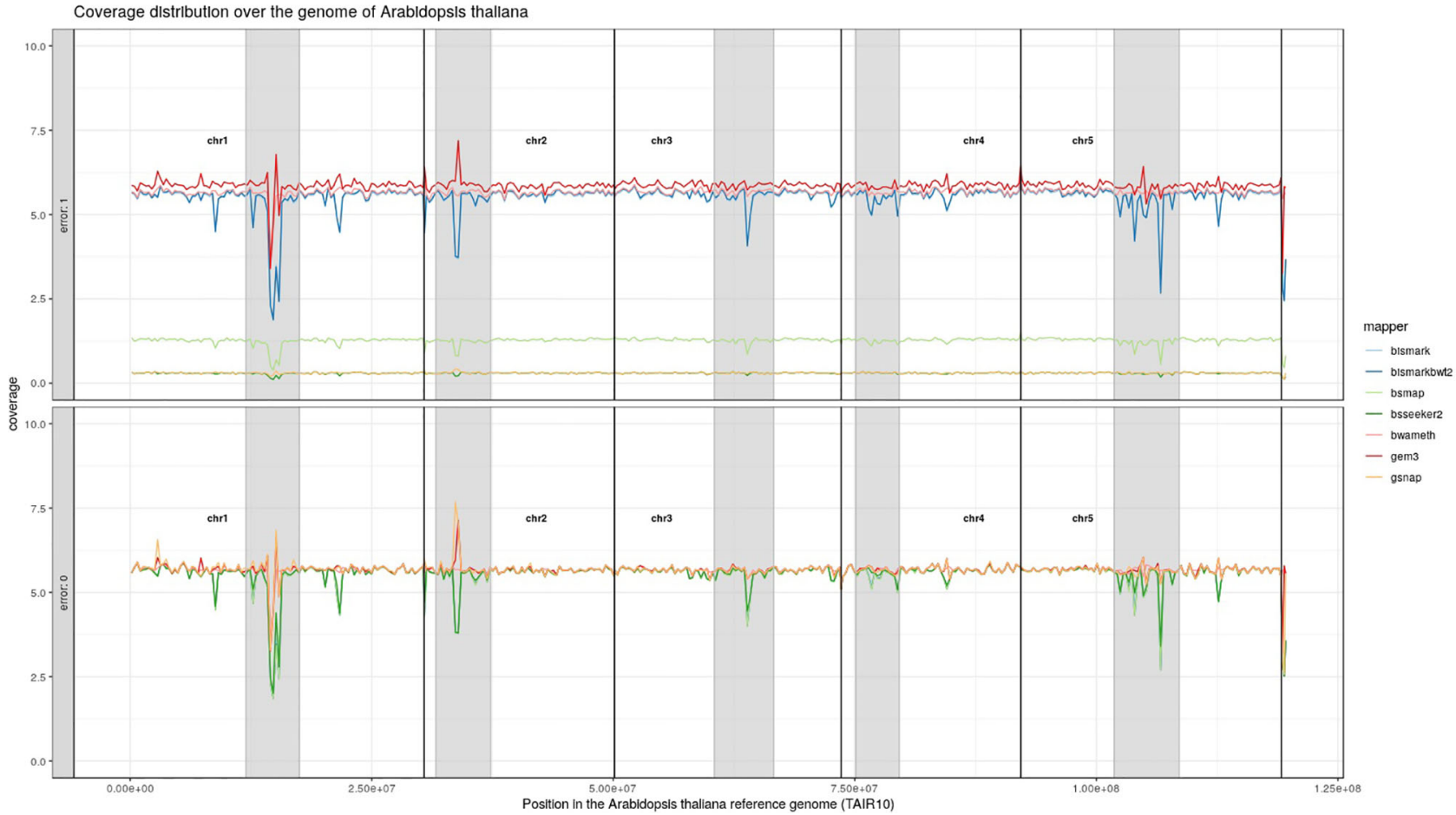

FIGURE 7 | Coverage distribution over the reference genome of *Arabidopsis thaliana* (TAIR10). The lower window shows the performance of 7 mapping tools using a simulated 5fold coverage dataset with 0% induced error rate, 100% conversion rate and 0 mismatches allowed. The upper window shows a simulated 5fold coverage dataset with an induced error rate of 1%, 100% conversion rate and with 0 mismatches allowed during the mapping. The number of reads has been calculated based on the ensemblPlants "Base Pairs" information. This could cause small differences to the estimated 5fold coverage datasets. Black lines indicate the borders of chromosomes. Grey regions highlight highly repetitive regions.

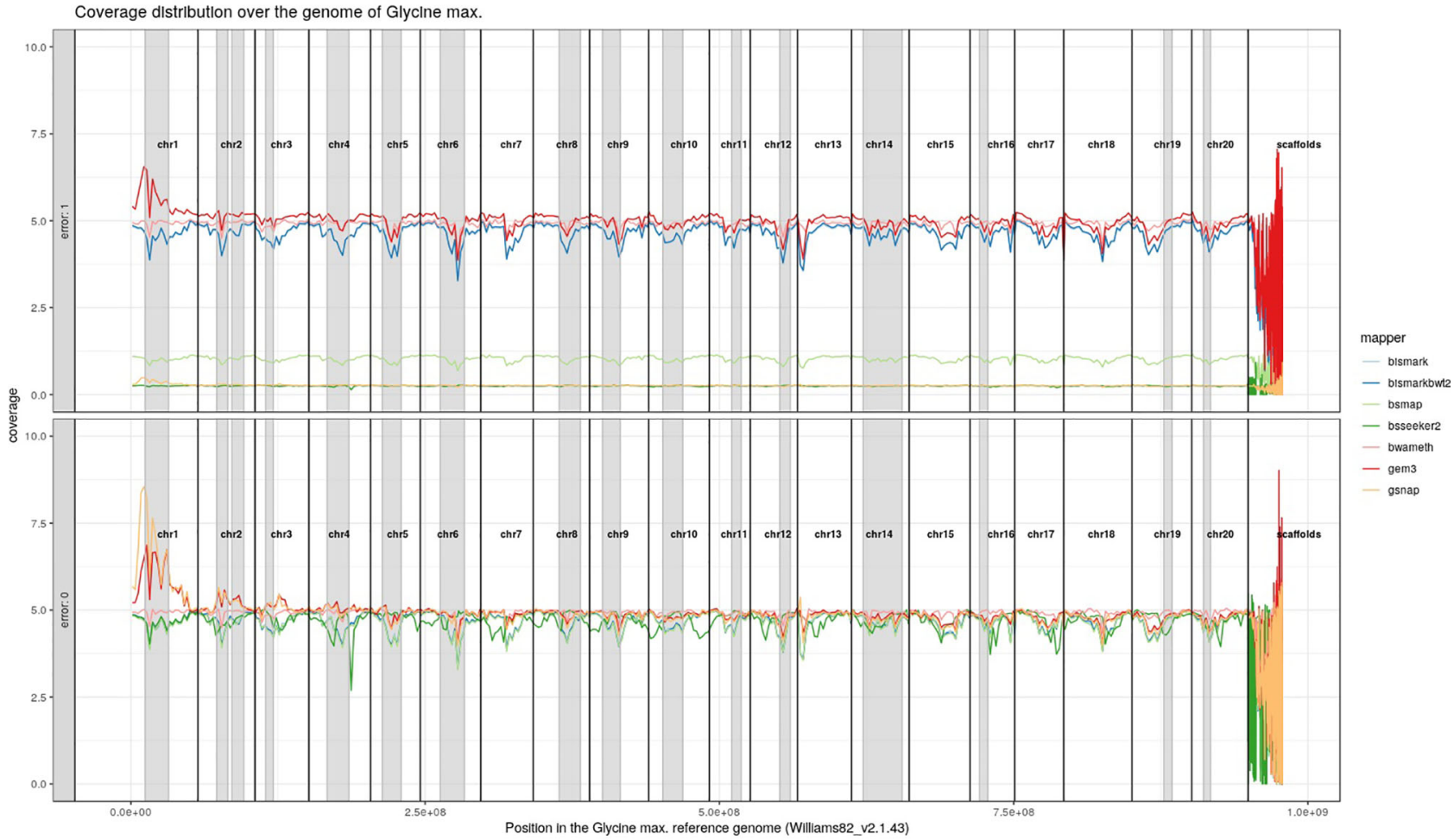

FIGURE 8 | Coverage distribution over the reference genome of *Glycine max* (Williams82_v2.1.43). The lower window shows the performance of 7 mappers using a simulated 5fold coverage dataset with 0% induced error rate, 100% conversion rate, and 0 mismatches allowed. The upper window shows a simulated 5fold coverage dataset with an induced error rate of 1%, 100% conversion rate, and with 0 mismatches allowed during the mapping. Black lines indicate the borders of chromosomes. Grey regions highlight highly repetitive regions.

TABLE 4 | Mean coverage of the four real data samples and the result of the DMR calling (SRR5044695 & SRR5044696 are the control and SRR5044699 & SRR5044700 are the heat stress replicates).

	SRR5044695	SRR5044696	SRR5044699	SRR5044700	DMRs
BismarkBwt2	15,0	14,5	17,9	19,2	281
Bwa-meth	32,5	29,5	35,6	38,8	256
GEM3	28,3	25,8	31,4	34,3	136
BismarkBwt1	11,3	11,0	13,9	14,7	97
GSNAP	10,3	9,5	12,0	12,6	70
BSMAP	10,5	9,9	12,2	12,7	63
BS-Seeker2	8,6	8,4	11,0	11,3	X

FIGURE 9 | Circular plot showing the distribution of DMRs (red) as result of mapping the same dataset with 6 different mapping tools on the Glycine max_v2.1 reference genome. The outer circle shows the chromosomes of *Glycine max* in blue. Blue lines indicate hypomethylation, whereas red lines indicate hypermethylation (see the full list of DMRs at https://github.com/grehl/benchWGBSmap). Numbers of overlapping DMRs could be found in the **Supplementary Material**.

mapped reads precisely but error rates above 0.5% and with 0 mismatches allowed during mapping leads to unique mapping rates below 25%. Other studies found Bismark to yield a reasonable combination of low memory consumption, low runtime, and high quality scores (Kunde-Ramamoorthy et al., 2014; Omony et al., 2019). This could be confirmed by our study, where Bismark showed the lowest memory consumption in all tested genomes. For runtime, we see high differences between Bismark using Bowtie and BismarkBwt2 under usage of Bowtie2. The precision showed good scores for all genomes and settings, with a slight decrease for the *Zea mays* genome.

For the second part of our study we mapped the same datasets with the seven mentioned mapping tools but had to exclude BS-Seeker2 for the DMR calling. Here, we see the most unique, proper paired reads for mapping with Bwa-meth and GEM3. Surprisingly, this could not be confirmed for the DMR calling where we obtained the most DMRs using Bismark with Bowtie2 using the same parameter sets, the same tools, and the same pipeline. We can only speculate what the reason for this behavior might be. Most likely this shift in the performance difference between the tools could be caused by false positive mappings which did not heavily influence the DMR calling, as they might have been mapped to "non-sense" positions either already involved in a DMR region, not causing much harm in remote regions due to the coverage threshold of 10-fold, or, they have been evenly distributed over treatment and control datasets.

CONCLUSION

In conclusion, we have shown high differences between the available mapping tools for bisulfite-treated reads based on simulated and real datasets in terms of runtime, memory consumption, and mapping quality. We see the stability of mapping quality against changes in the conversion rate, high differences between the mapping tools in terms of the number of uniquely mapped reads as well as in the capability to map correctly under the impact of higher error rates in five different genomes. Additionally, we see high differences with regard to the analyzed genome, dependent on the size and structure of repeats.

For *Arabidopsis thaliana* basically every one of the examined mapping tools could be used with a sufficient mapping rate and good quality, at least when assuming a low error rate. This holds true for low error rates in *Glycine max* mappings. For higher error rates we recommend Bwa-meth as well as Bismark, using Bowtie1 or Bowtie2. For paralogue-rich species such as *Brassica napus*, polyploid species such as *Solanum tuberosum*, or large genomes with many repetitive sequences such as in *Zea mays* we prefer correct mappings over a large number of unique mapped reads. Therefore, going with Bismark using Bowtie1 or Bowtie2 or BSMAP and BS-Seeker2 with a higher number of mismatches allowed might work well, looking at the perspective of mapping.

Altogether, we recommend BSMAP as this requires the shortest run time and yields the highest precision and Bismark which requires the smallest amount of memory and yields high precision and high numbers of uniquely mapped reads. Furthermore, Bwa-meth could be used with care in terms of precise calling of DMRs.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

Conceptualization: CG and MW Methodology: CG, MW Investigation: CG, MW. Writing—original draft preparation: CG and IL. Writing—review and editing: IL, BG, IG. Visualization: CG and MW. Supervision: BG, IG. Project administration: CG, BG, IG. Funding acquisition: CG, BG, IG.

FUNDING

This research was funded by the state of Saxony-Anhalt and Volkswagen Stiftung. The Article Processing Charges have been supported by the open-science fund of the Martin-Luther University Halle-Wittenberg. IL was funded by the Deutsche Forschungsgemeinschaft (DFG, Germany) Research Training Group 2155 (ProMoAge).

ACKNOWLEDGMENTS

We thank all the members of the Bioinformatics Group of Halle-Wittenberg University, especially Alexander Gabel, Jan Grau, Claus Weinholdt, Silvio Weging, and Samar Fatma for valuable discussions as well as for support in terms of biostatistics and visualization.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00176/full#supplementary-material

REFERENCES

Aberg, K. A., Chan, R. F., Shabalin, A. A., Zhao, M., Turecki, G., Staunstrup, N. H., et al. (2017). A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics* 12 (9), 743–750. doi: 10.1080/15592294.2017.1335849

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., et al. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* 28 (10), 1106–1114. doi: 10.1038/nbt.1681

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Sci. (New York N.Y.)* 345 (6199), 950–953. doi: 10.1126/science.1253435

Chen, P.-Y., Cokus, S. J., and Pellegrini, M. (2010). BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinf.* 11, 203. doi: 10.1186/1471-2105-11-203

Chen, X., Schönberger, B., Menz, J., and Ludewig, U. (2018). Plasticity of DNA methylation and gene expression under zinc deficiency in Arabidopsis roots. *Plant Cell Physiol.* 59 (9), 1790–1802 doi: 10.1093/pcp/pcy100

Condon, D. E., Tran, P. V., Lien, Y.-C., Schug, J., Georgieff, M. K., Simmons, R. A., et al. (2018). Defiant: (DMRs: easy, fast, identification and ANnoTation) identifies differentially Methylated regions from iron-deficient rat hippocampus. *BMC Bioinf.* 19 (1), 31. doi: 10.1186/s12859-018-2037-1

Escalona, M., Rocha, S., and Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* 17 (8), 459–469. doi: 10.1038/nrg.2016.57

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS One* 6 (1), e16526. doi: 10.1371/journal.pone.0016526

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinf. (Oxford England)* 28 (20), 2678–2679. doi: 10.1093/bioinformatics/bts503

Grehl, C., Kuhlmann, M., Becker, C., Glaser, B., and Grosse, I. (2018). How to design a whole-genome Bisulfite sequencing experiment. *Epigenomes* 2 (4), 21. doi: 10.3390/epigenomes2040021

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., et al. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* 14, 774. doi: 10.1186/1471-2164-14-774

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., et al. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PloS Comput. Biol.* 5 (9), e1000502. doi: 10.1371/journal.pcbi.1000502

Hossain, M. S., Kawakatsu, T., Kim, K. D., Zhang, N., Nguyen, C. T., Khan, S. M., et al. (2017). Divergent cytosine DNA methylation patterns in single-cell, soybean root hairs. *New Phytol.* 214 (2), 808–819. doi: 10.1111/nph.14421

Köster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinf. (Oxford England)* 28 (19), 2520–2522. doi: 10.1093/bioinformatics/bts480

Koch, A., Joosten, S. C., Feng, Z., de Ruijter, T. C., Draht, M. X., Melotte, V., et al. (2018). Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* 15 (7), 459–466. doi: 10.1038/s41571-018-0004-4

Krueger, F., and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-seq applications. *Bioinf. (Oxford England)* 27 (11), 1571–1572. doi: 10.1093/bioinformatics/btr167

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N. J., Harris, R. A., Xu, M., et al. (2014). Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res.* 42 (6), e43. doi: 10.1093/nar/gkt1325

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* 11 (3), 191–203. doi: 10.1038/nrg2732

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40 ((Database issue)), D1202–D1210. doi: 10.1093/nar/gkr1090

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10 (3), R25. doi: 10.1186/gb-2009-10-3-r25

Lauss, K., Wardenaar, R., Oka, R., van Hulten, M. H. A., Guryev, V., Keurentjes, J. J. B., et al. (2018). Parental DNA Methylation states are associated with heterosis in epigenetic hybrids. *Plant Physiol.* 176 (2), 1627–1645. doi: 10.1104/pp.17.01054

Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., and Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front. In Genet.* 4, 288. doi: 10.3389/fgene.2013.00288

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinf. (Oxford England)* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinf. (Oxford England)* 26 (5), 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinf. (Oxford England)* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

S. Liu, R. Snowdon and B. Chalhoub (Eds.) (2018). *The Brassica napus Genome* (Cham: Springer International Publishing (Compendium of Plant Genomes)). doi: 10.1007/978-3-319-43694-4

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* 9 (12), 1185–1188. doi: 10.1038/nmeth.2221

Mehra, M., Gangwar, I., Shankar, R., and Houben, A. (2015). A deluge of complex repeats. The solanum genome. *PloS One* 10 (8), e0133962. doi: 10.1371/journal.pone.0133962

Omony, J., Nussbaumer, T., and Gutzat, R. (2019). DNA methylation analysis in plants. Review of computational tools and future perspectives. *Briefings In Bioinf.* 38 (5), 285. doi: 10.1093/bib/bbz039

Otto, C., Stadler, P. F., and Hoffmann, S. (2012). Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinf. (Oxford England)* 28 (13), 1698–1704. doi: 10.1093/bioinformatics/bts254

Pedersen, B., Eyring, K., De, S., Yang, I., and Schwartz, D. (2014). Fast and accurate alignment of long bisulfite-seq reads. Prepring: arXiv:1401.1129v2. *Bioinf. (Oxford England).* 5/13/2014.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463 (7278), 178–183. doi: 10.1038/nature08670

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Sci. (New York N.Y.)* 326 (5956), 1112–1115. doi: 10.1126/science.1178534

Sherman-Broyles, S., Bombarely, A., Grimwood, J., Schmutz, J., and Doyle, J. (2014). Complete plastome sequences from Glycine syndetika and six additional perennial wild relatives of soybean. *G3 (Bethesda Md.)* 4 (10), 2023–2033. doi: 10.1534/g3.114.012690

Sun, Z., Cunningham, J., Slager, S., and Kocher, J.-P. (2015). Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* 7 (5), 813–828. doi: 10.2217/epi.15.21

Sun, X., Han, Y., Zhou, L., Chen, E., Lu, B., Liu, Y., et al. (2018). A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinf. (Oxford England)* 34 (16), 2715–2723. doi: 10.1093/bioinformatics/bty174

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinf. (Oxford England)* 31 (12), 2032–2034. doi: 10.1093/bioinformatics/btv098

Tran, H., Porter, J., Sun, M.-A., Xie, H., and Zhang, L. (2014). Objective and comprehensive evaluation of bisulfite short read mapping tools. *Adv. In Bioinf.* 2014, 472045. doi: 10.1155/2014/472045

Wolf, P. G., Sessa, E. B., Marchant, D. B., Li, F.-W., Rothfels, C. J., Sigel, E. M., et al. (2015). An exploration into fern genome space. *Genome Biol. Evol.* 7 (9), 2533–2544. doi: 10.1093/gbe/evv163

Wu, T. D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinf. (Oxford England)* 26 (7), 873–881. doi: 10.1093/bioinformatics/btq057

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016). GMAP and GSNAP for genomic sequence alignment: enhancements to speed,

accuracy, and functionality. *Methods In Mol. Biol. (Clifton N.J.)* 1418, 283–334. doi: 10.1007/978-1-4939-3578-9_15

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf.* 10, 232. doi: 10.1186/1471-2105-10-232

Xie, M., Chung, C. Y.-L., Li, M.-W., Wong, F.-L., Wang, X., Liu, A., et al. (2019). A reference-grade wild soybean genome. *Nat. Commun.* 10 (1), 1216. doi: 10.1038/s41467-019-09142-9

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., et al. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475 (7355), 189–195. doi: 10.1038/nature10158

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 Plant Science

ORIGINAL RESEARCH
published: 04 March 2020
doi: 10.3389/fpls.2020.00156

A Partially Phase-Separated Genome Sequence Assembly of the *Vitis* Rootstock 'Börner' (*Vitis riparia* × *Vitis cinerea*) and Its Exploitation for Marker Development and Targeted Mapping

*Daniela Holtgräwe[1], Thomas Rosleff Soerensen[1], Ludger Hausmann[2], Boas Pucker[1], Prisca Viehöver[1], Reinhard Töpfer[2] and Bernd Weisshaar[1]**

[1] Faculty of Biology, Center for Biotechnology, Bielefeld University, Bielefeld, Germany, [2] Institute for Grapevine Breeding Geilweilerhof, Julius Kuehn-Institute, Federal Research Centre for Cultivated Plants, Siebeldingen, Germany

OPEN ACCESS

Edited by:
Uwe Scholz, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
Cheng Zou, Cornell University, United States
Surya Sapkota, Cornell Tech, United States

***Correspondence:**
Bernd Weisshaar
bernd.weisshaar@uni-bielefeld.de

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *25 October 2019*
Accepted: *31 January 2020*
Published: *04 March 2020*

Citation:
Holtgräwe D, Rosleff Soerensen T, Hausmann L, Pucker B, Viehöver P, Töpfer R and Weisshaar B (2020) A Partially Phase-Separated Genome Sequence Assembly of the Vitis Rootstock 'Börner' (Vitis riparia × Vitis cinerea) and Its Exploitation for Marker Development and Targeted Mapping. Front. Plant Sci. 11:156. doi: 10.3389/fpls.2020.00156

Grapevine breeding has become highly relevant due to upcoming challenges like climate change, a decrease in the number of available fungicides, increasing public concern about plant protection, and the demand for a sustainable production. Downy mildew caused by *Plasmopara viticola* is one of the most devastating diseases worldwide of cultivated *Vitis vinifera*. In modern breeding programs, therefore, genetic marker technologies and genomic data are used to develop new cultivars with defined and stacked resistance loci. Potential sources of resistance are wild species of American or Asian origin. The interspecific hybrid of *Vitis riparia* Gm 183 x *Vitis cinerea* Arnold, available as the rootstock cultivar 'Börner,' carries several relevant resistance loci. We applied next-generation sequencing to enable the reliable identification of simple sequence repeats (SSR), and we also generated a draft genome sequence assembly of 'Börner' to access genome-wide sequence variations in a comprehensive and highly reliable way. These data were used to cover the 'Börner' genome with genetic marker positions. A subset of these marker positions was used for targeted mapping of the *P. viticola* resistance locus, *Rpv14*, to validate the marker position list. Based on the reference genome sequence PN40024, the position of this resistance locus can be narrowed down to less than 0.5 Mbp on chromosome 5.

Keywords: *de novo* genome assembly, *Vitis*, rootstock, targeted mapping, SSR marker, SNV detection, whole genome sequencing

INTRODUCTION

Downy mildew of grapevine is caused by the oomycete *Plasmopara viticola*. It is a serious threat for viticulture especially in humid and warm environments. Intensive chemical plant protection, frequently through the use of copper, is necessary to protect grapevines under disease-promoting conditions. Resistance breeding allows for the selection of new cultivars that require a reduced

number of chemical treatments. Genetic analyses of resistance loci have promoted such breeding initiatives that are today oriented towards stacking (combining) of several resistance loci (Töpfer and Eibach, 2016). In the recent past, therefore, a number of resistance loci for downy mildew have been identified, and these have been localized within grapevine genomes. Around two dozen loci designated *Rpv* (Resistance *Plasmopara viticola*) are listed in the Vitis International Variety Catalogue [VIVC, www.vivc.de; and (Hausmann et al., 2019)]. Some of them are of elite genetic background, such as *Rpv1* (Merdinoglu et al., 2003), *Rpv3* (Welter et al., 2007; Bellin et al., 2009), *Rpv10* (Schwander et al., 2012), and *Rpv12* (Venuti et al., 2013), and these can be used in breeding programs for selection of new cultivars. Other loci are less well characterized and are essentially known only from wild species (Marguerit et al., 2009; Blasi et al., 2011; Ochssner et al., 2016). To allow their application, these loci need to be introgressed into a favorable *Vitis vinifera* genetic background. Such an introgression can be accelerated considerably by using marker-assisted selection or marker-assisted back crossing (Herzog et al., 2013). To start such an approach, a locus should be well characterized, and sufficient closely linked markers should be available. Within plant breeding programs, simple sequence repeats (SSRs) and single nucleotide variant (SNVs) markers are widely applied. The discovery and development of this marker types by using NGS data has been successfully demonstrated for various plant species (Taheri et al., 2018).

Since the grapevine reference genome sequence became available (Jaillon et al., 2007), QTL mapping by using SSR markers has become simplified in a reference sequence-guided approach [e.g. (Schwander et al., 2012)]. Alternatively, next-generation sequencing approaches (NGS) were successfully used to speed up mapping processes (Barba et al., 2014; Hyma et al., 2015; Capistrano-Gossmann et al., 2017).

One of the genetically useful genotypes as a source for resistance is the rootstock cultivar 'Börner'. It is an interspecific hybrid of *V. riparia* Gm 183 x *V. cinerea* Arnold; 'Börner' is thus derived from individuals of two *Vitis* species native to North America that carry several resistances. Among these are resistances to phylloxera [*Rdv1* (Zhang et al., 2009)], black rot [*Rgb1* and *Rgb2* (Rex et al., 2014)], and to downy mildew [*Rpv14* (Ochssner et al., 2016)]. To support the molecular analysis of these and other loci, we generated a first draft genome sequence assembly of the highly heterozygous 'Börner' genome. We produced comprehensive lists of SSR and SNV marker positions based on NGS data from 'Börner,' and we validated these marker positions randomly as well as by using *Rpv14* on chromosome 5 (chr.5) as an example target.

The first grapevine reference genome sequence was generated based on Sanger sequencing technology using the 'Pinot noir' inbreeding line PN40024, which is estimated to be homozygous for 93% of the genome, thus enabling a successful assembly [(Jaillon et al., 2007), current version 12x.v2]. However, PN40024 suffers from inbreeding depression, and selfing grapevine for homozygosity is a difficult and a time-consuming process. Up until 2018, other available grapevine genome sequences were derived from heterozygous genotypes (Velasco et al., 2007; Adam-Blondon et al., 2011; Di Genova et al., 2014), and this posed a challenging sequencing and assembly problem that is not yet fully solved. In 2007, it was impossible to achieve a high-quality genome assembly for 'Pinot noir' due to its heterozygosity since merged allelic sequences made it impossible to separate both haplophases. As NGS techniques developed progressively, the number of sequenced plant genomes has rapidly increased, breaching the 100 mark some time ago (Michael and Jackson, 2013; VanBuren et al., 2015). Only a few genome sequences are based purely on the Sanger sequencing technique (e.g., *Arabidopsis* and grapevine). A larger number was generated based on mixed data types, including Sanger, 454, and Illumina (e.g., potato, sugar beet, and *Brassica rapa*) or only Illumina-derived sequences (e.g., eggplant and Norway spruce). The quality of the latter, generated by early NGS technologies, does not reach the quality of the older BAC-based Sanger assemblies despite higher read data coverage (Alkan et al., 2011; Burton et al., 2013). Although long reads became the best option for generating high-quality genome sequences (Li et al., 2018), the high throughput and low costs of Illumina sequencing has made it economical to sequence plenty of genomes at least to a draft state or to re-sequence individual genotypes. A major problem and current challenge in plant genome sequence assembly is resolving the repeat structures and proper continuous separation of haplophases within heterozygous or polyploid plants (VanBuren et al., 2015). Repeat sequences are very common in plant genomes, including repeats that are more than 10 kbp in length. Therefore, resolving plant genome complexity with short NGS reads, mainly 100–300 nt in length, is a big challenge. Such assemblies result in collapsed sequences and breakdown at the borders of repetitive regions, and this causes shortening of contigs. Some improvements were reached by using a variation of paired-end libraries (VanBuren et al., 2015) combined with mate-pair libraries. Long-insert libraries were of particular value to bridge such gaps up to about 12 kbp.

Here, we utilized sequencing data mainly generated by 454 and Illumina technologies to generate a draft assembly of the 'Börner' genome sequence with the main goal of allowing easy access to markers. The approach was complemented by sequencing BAC ends from a library created from 'Börner' DNA. We hypothesized that the interspecific nature of 'Börner' and the genetic distance of its two wild ancestor species, *V. riparia* and *V. cinerea*, could be helpful to dissect and assign the haplophases of 'Börner' in its parental genomes even with Illumina reads. Having at least to a certain extent phased contigs at hand, several downstream applications become feasible. The alignment of 'Börner' contigs to the reference PN40024 sequence revealed many polymorphic sites, including those at SSR positions, as being sources of potential high value for, to give an example, the purpose of grapevine resistance breeding. The detection of single SNVs or groups of nearby SNVs within a multiple alignment of two 'Börner' contigs and the PN40024 reference sequence are both considered to be useful in a bulked segregant analysis (BSA). The identification of

markers linked to loci or genes causing disease resistance by applying BSA has been shown in plants (Michelmore et al., 1991). Pooling equal amounts of DNA is a practical way to reduce the cost of large-scale association studies to identify susceptibility loci for diseases (Sham et al., 2002; Zou et al., 2016). The pooling allows us to measure allele frequencies in groups of individuals by using far fewer genotyping assays than required when genotyping all individuals. Here, we demonstrate the power of heterozygous positions within one species to physical narrow down a candidate locus for downy mildew resistance (*Rpv14*).

MATERIAL AND METHODS

BAC Library

To construct a BAC library of the grapevine cv. 'Börner' young frozen leaf material was sent to Bio S&T Inc. (Montreal, Canada) as a service provider for the generation of BAC libraries. The DNA was inserted into the vector pIndigoBAC-5 (Epicentre, Madison, USA; ENA/GenBank accession number EU140754) cut with *Hin*dIII. Competent cells of *Escherichia coli* strain DH10B were transformed with the ligation mix and plated on LB medium supplemented with 12.5 µg/mL chloramphenicol (CM), 40 µg/mL X-GAL and 0.2 mM IPTG. White colonies were picked and transferred in 384-well microtiter plates filled with LB freezing medium (Zimmer and Verrinder Gibbins, 1997) containing CM. Altogether, 159 microtiter plates with 384 wells were generated, and these were duplicated and stored at -80°C. The insert size was estimated by pulsed-field gel electrophoresis to be on average 93 kbp. The library coverage is almost 12 haploid genome equivalents based on a genome size of about 500 Mbp (Lodhi and Reisch, 1995).

BAC End Sequencing by Sanger Sequencing

A set of 43,008 (112 x 384) clones, comprising 8.4 haploid genome equivalents, of the 'Börner' BAC library (IGS-SCF-1083) was cultivated, harvested, and applied to DNA extraction and subsequently sequenced from both ends on an ABI3730XL DNA analyzer (Applied Biosystems, Foster City, California, USA) as described before (Dohm et al., 2012). Sequencing primers were ccfw and ccrv for plate 1–52 and 55. Plates 53, 54, and 56–112 were sequenced using ccrv and a modified sequencing primer, pIfw (**Supplementary Table 1**).

Processing of BAC End Sequences (BES)

Trace files were processed by the base calling program phred (www.phrap.org, version 0.071220.c). Vector sequences were trimmed using the program blastall (Altschul et al., 1990) in version 2.2.24 with blast parameters set to the ones suggested by NCBI (www.ncbi.nlm.nih.gov/tools/vecscreen/about/). Low-quality sequences were masked by applying a sliding window approach (size: 10 bp, minimal average score > 13), and the longest unmasked subsequence was taken as a high-quality sequence if the length was ≥ 80 bp. Slippage reads were removed as described before (Dohm et al., 2012). The dataset was filtered for contamination with *E. coli* sequences (blastall matches with an e-value ≤ 1e-50). Afterwards, the sequences were singularized per BAC end. A total of 69,444 high-quality 'Börner' BES were submitted to ENA/GenBank and can be accessed under accession numbers KG622866 - KG692309.

NGS Sequencing Library Preparation

'Börner' DNA, extracted from young leaf material with a commercial kit (Qiagen) according to the instructions of the manufacturer, was sequenced using 454 GS-FLX Titanium (Roche), IonTorrent PGM (life technologies), Illumina GAIIx, MiSeq, and HiSeq1500 instrumentation. The preparation of Illumina paired-end (PE) and single-end (SE) sequencing libraries was performed according to the Illumina TruSeq DNA Sample Preparation v2 Guide. DNA was fragmented by nebulization. After end repair and A-tailing, individual indexed adaptors were ligated to the DNA fragments, thus allowing for multiplexed sequencing. The adaptor-ligated fragments were size selected on a 2% low melt agarose gel to a size of 300–600 bp. After enrichment PCR of fragments that carry adaptors on both ends, the final libraries were quantified with PicoGreen. The average fragment size of each library was determined on a BioAnalyzer High-Sensitivity DNA chip. Libraries were sequenced either 100 nt on an SE run using one lane of a GAIIx run (ENA/SRA accession numbers ERR2729619 and ERR2729620) or sequenced 2 x 100 nt PE on GAIIx (ERR2729739, ERR2729740, ERR2729741, ERR2729743, and ERR2729744) or HiSeq1500 (ERR2729742) sequencers. A mate-pair library was constructed according to the Illumina Mate Pair Sample Preparation v2 Guide. DNA was fragmented to a size of 3–4 kbp by using Hydroshear. Fragment ends were repaired and biotinylated. Afterwards, fragments of 3.5 kbp size were selected by gel electrophoresis. The resulting molecules were circularized. Removal of linear fragments was achieved by DNA exonuclease treatment. The circular molecules were randomly fragmented by nebulization. The biotinylated fragments were purified using streptavidin-coated magnetic beads. Several steps were carried out as described for the PE library construction. Finally, the library was sequenced in paired-end modus using an Illumina MiSeq generating 2 x 36 nt PE reads (ERR2729745).

GS-FLX sequencing libraries (454) were prepared from 10 micrograms of DNA according to the GS-FLX-Titanium Rapid Library Preparation Method Manual. DNA was fragmented by nebulization. Purified DNA fragments were subjected to end repair and subsequently to adaptor ligation. Large DNA fragments were selected *via* AMPure Beads. After emPCR and enrichment of DNA carrying beads, sequencing was performed on several runs on a Roche/FLX sequencing instrument using Titanium XLR70 sequencing kits (ERR2728011-ERR2718029). Mate-pair libraries (ERR2728288-ERR2728292) with varying insert length (multi-span PE reads, 3 kbp, and 8 kbp) were prepared from 15–20 micrograms of genomic DNA according to GS-FLX-Titanium manual method and using the GS FLX Titanium Paired End Adaptors kit and sequenced on a Roche/FLX sequencer.

The library construction for IonTorrent PGM (life technologies) sequencing was performed according to the Ion Xpress™ Plus gDNA and Amplicon Library Preparation guide for a 200 bp library. DNA was fragmented by nebulization. The ends were repaired and ligated to adapters. Size selection of DNA molecules was performed *via* E-Gel® SizeSelect™ Agarose Gel to an average insert size of 350 bp. After enrichment of the template carrying Ion Sphere particles on an Ion OneTouch, sequencing was performed using a 318 chip on an Ion Personal Genome Machine™ system (ERR2729812).

In addition to 'Börner' we used the *V. vinifera* cultivar 'Pinot noir' for DNA extraction, library construction, and Illumina HiSeq1500 sequencing (2 x 100 nt PE) according to the workflow described above. The read data were submitted to the ENA study PRJEB36326 with the read accessions ERR3834961 and ERR3834962).

NGS Read Pre-Processing

Illumina reads were quality trimmed and adapter clipped using Trimmomatic (Bolger et al., 2014) with all available adapter sequences from Illumina. Reads from the 454- and PGM/Ion Torrent-sequencing platform were trimmed with the CLC Genomics Workbench using the trim points of the SFF-files. Reads shorter than 36 nt were discarded. After pre-processing, 75.62 Gbp raw data (=90.9% of the raw bases) were available for *de novo* assembly.

Assembly, Post-Processing, and Validation

A *de novo* hybrid assembly of the 'Börner' genome was generated with the CLC Genomics Workbench (v. 8.0) using the certain parameters: length fraction 0.5, similarity fraction 0.8, word size 20, and bubble size 50. For the re-mapping of the reads to the contigs that the CLC assembler carries out, the filtering thresholds were optimized to a length fraction of 0.8 (default: 0.5) and a similarity fraction of 0.95 (default: 0.9). From 75.62 Gbp of read data, representing nearly 80-fold coverage for each of the two 'Börner' haplophases, 338,484 contigs with a total length of 652.17 Mbp were calculated. A total of 91.3% of the reads could be mapped back to the assembly, indicating a high probability that the sequence data was integrated correctly into the contigs. This raw assembly was designated BoeWGS0.8.1 (**Supplementary Table 2**).

The assembly metadata include a value for the number of reads that cover a given position and averages of these values for each contig. The value, which is essentially the total length of aligned reads divided by the assembly size, is referred to below as "alignment depth" (AD) and was used to deduce if a given contig represents a single haplophase or contains merged sequence information from both haplophases.

To test the completeness of the assembly we mapped the 69,444 'Börner' BES on BoeWGS0.8.1. Mapping was done with the program BLAT (Kent, 2002). The majority (43,008 BES = 61.9%) mapped with 99% identity or better and for a length fraction of 0.5. About 4.5% BES matched the end of contigs with between 20 to 50% of their length, but their mapping was not unique to one contig within the assembly. A small proportion of these (733 BES) produced multiple matches, suggesting that these contain mainly repetitive sequences. To avoid connecting contigs from different phases, we refrained from scaffolding since we had no information about the phase from which the BACs/BES were derived.

The raw assembly BoeWGS0.8.1 was checked for contaminations by searching against a custom database with sequences of human, yeast, mold, several fungal and viral pathogens, and insects, such as aphids and mites, with blast+ (e-value <=1e-50) (Camacho et al., 2009). A BLAST (Altschul et al., 1990) run querying GenBank's non-redundant protein database (nr) completed the filtering process by using the same cut off criteria. Contigs matching to contaminants were also characterized by very low AD (1–20x). The removal of PhiX as well as sequences resembling cloning vectors caused a reduction of 213,834 bp and a total removal of three contigs. Plastid sequences from *V. vinifera* were detected in 67 contigs and removed as well. In addition, contigs smaller than 500 bp were discarded. Although nearly 30% of all initial contigs were rejected by this filter, the total assembly size was only reduced by about 5%. A total of 238,193 contigs remained after the different filters and were included in the final assembly, which was designated BoeWGS1.0 (Biosample PRJEB28084/http://www.ebi.ac.uk/ena/data/view/PRJEB28084). The results were evaluated for standard assembly statistics parameters (**Supplementary Table 2**).

The BoeWGS1.0 contigs were aligned to the reference genome sequence PN40024 12x.v2 (https://urgi.versailles.inra.fr/download/vitis/12Xv2_grapevine_genome_assembly.fa.gz) with blast+ with the dust filter turned off and using an e-value cutoff of 1e-50 and a culling limit of 1. As a result of this analysis, each mapping contig from BoeWGS1.0 receives location information derived from PN40024. If exactly two BoeWGS1.0 contigs were assigned to one locus of the reference sequence, we took this as indication that the assembly separated the two 'Börner' haplophases. We have referred to these contigs, which may contain allelic sequence information, as "contig pairs".

BWA-MEM (Li, 2013) was applied to map all PE Illumina reads to the final assembly sequence using the –M option to mark short splits as secondary. The resulting mapping was subjected to inspection by REAPR (Hunt et al., 2013) to assess the assembly quality based on coverage, distance of PE reads, and orientation of PE reads.

BUSCO v3 (Simão et al., 2015) was deployed to analyze the recovery of the 'embryophyta odb9' benchmarking sequences in the assembly. This analysis was performed in the 'genome' mode using an e-value cutoff of 1e-3 and considering up to three candidate regions per BUSCO hit.

Calculation of Heterozygosity

Jellyfish v2.2.4 (Marcais and Kingsford, 2011) was applied on the reads of ERR2729742 to calculate histograms for k-mer lengths 19, 21, 23, and 25. The results were individually subjected to GenomeScope (Vurture et al., 2017) using default parameters for calculation of the heterozygosity. The average of all four resulting predictions was calculated to avoid k-mer length dependency of the final result. For comparison, we included Illumina read data from 'Pinot noir' (*V. vinifera*). As for the analysis of 'Börner,'

about 40 Gbp Illumina PE data were used and the analyses were performed with the same parameters.

Genome-Wide Variant Detection

A strict mapping of 245,325,650 'Börner' PE reads (2x100 nt) on the PN40024 12x.v2 reference sequence with an average coverage of 50x formed the basis for the variant detection. The mapping was done with a CLC Genomics Workbench toolkit, the similarity fraction set to 0.95 and the length fraction to 0.9. Non-specific matches were discarded. Variants [single nucleotide variants (SNVs), insertions and deletions (InDels), as well as multi nucleotide variations (MNVs)] were detected with CLC's Probabilistic Variant Caller using default parameters. Only bi-allelic SNVs with a read coverage between 10 and 90 were kept for downstream analyses, whereas tri-allelic and poly-allelic variants, InDels, and MNVs were ignored.

Bi-allelic SNVs were classified as homozygous variants if both 'Börner' haplotypes differed from the reference sequence (in these cases both 'Börner' haplophases contain the same nucleotide at the inspected position), whereas heterozygous bi-allelic SNVs were those where only one 'Börner' haplotype differred from the reference sequence. The complete SNV data is available at https://doi.org/10.4119/unibi/2938178 and https://doi.org/10.4119/unibi/2938180.

Simple Sequence Repeat Detection

SSR detection was done with MISA [MIcroSAtellite identification tool; http://pgrc.ipk-gatersleben.de/misa/misa.html, (Thiel et al., 2003)] using default parameters. Possible SSR marker positions were evaluated by comparison of the candidate SSRs plus adjacent +/-200 bases from BoeWGS1.0 with the SSRs of the PN40024 12x.v2 reference sequence. The comparison was performed with blastn with dust filter turned off and 1e-50 as e-value cutoff. SSR candidates where a BoeWGS1.0 contig pair matched one coordinate in the reference sequence and differences did not exceed 40 bp (identity >=90%) were classified as useful for SSR marker development. The SSR candidates were named "SSR [chr#]_[contig-ID]" with a trailing number if more than one SSR candidate was detected for one contig pair.

SSR Marker Validation

For the validation of heterozygous SSRs within the BoeWGS1.0 assembly and variants in relation to the reference sequence, 53 SSR candidate positions were randomly selected for SSR marker validation. The SSRs from different genomic regions were selected with a minimum of two nucleotides and of six contiguous repeat units in the reference sequence. Primer3 (Untergasser et al., 2012) was used to design the amplimers (two primers directing towards each other based on the source sequence) fitting to the BoeWGS1.0 contig sequences as well as to the reference sequence PN40024 (**Supplementary Table 1**). The expected amplicon size of the markers was set from 150 to 400 bp and the primer size from 18 to 27 nt. DNA from 'Börner' and V3125 (see below) was used to check the polymorphism of the SSR marker sequences by PCR and gel electrophoresis on a 3% agarose gel. A few PCR products with expected smaller sequence variations were applied to Sanger sequencing and analyzed at the bp level by multiple sequence alignments. Verified heterozygous markers either in 'Börner' and/or in relation to V3125 were analyzed by fragment analysis using DNA of the crossing parents (V3125 and 'Börner') as well as the parents of 'Börner' as described before (Ochssner et al., 2016)

Locus-Guided Variant Calling and Marker Generation

The BoeWGS1.0 contig mapping on the PN40024 12x.v2 reference sequence was evaluated for a genomic region on chr.5. The BLAST results from the post-processing analysis described above were used. Regions where a multiple alignment composed of exactly two BoeWGS1.0 contigs (i.e., contig pairs) and parts of chr.5 from the PN40024 12x.v2 reference sequence was built were considered as genome regions representing both haplophases if the contig pairs selected showed the expected AD of about 75x (75 ± 25).

Each contig pair was aligned with edialign (from the EMBOSS suite, version 6.2) using default parameters. Alignments without N-stretches and with lengths of at least 500 bp were subjected to SNV detection. The list of contig pairs used for SNV detection on chr.5 is available as **Supplementary Table 3**. The downstream analysis focused on grouped SNVs between BoeWGS1.0 contig pairs that are bi-allelic after comparison with the PN40024 reference sequence.

Amplimer Design for the *Rpv14* Locus

BoeWGS1.0 contig pairs on chr.5 were used for SNV-based marker development addressing the *Rpv14* locus. Amplimers were designed using Primer3 (Untergasser et al., 2012) preferably for contig pair alignments with a rate between 0.4 and 1 SNVs per 100 bp (**Supplementary Table 1** and **Supplementary Table 3**). We targeted an expected amplicon size between 400 and 530 bp. Deduced amplimers were tested to be unique in BoeWGS1.0 and PN40024 12x.v2.

Plant Material, Bulk Set Up, DNA Extraction, and PCR

A biparental F1 mapping population comprising 202 individuals of the grapevine breeding line V3125 ('Schiava Grossa' x 'Riesling') and 'Börner' (*V. riparia* Geisenheim 183 x *V. cinerea* Arnold), cultivated in the field at the JKI Institute for Grapevine Breeding Geilweilerhof, Siebeldingen, Germany, has been phenotyped three times for downy mildew leaf resistance as described previously (Ochssner et al., 2016). For SNV genotyping by amplicon sequencing, genomic DNA was extracted from the parents V3125, and 'Börner' and selected F1 genotypes of the mapping population. Building bulks of DNA from all the individual samples and typing one SNV marker at a time can save valuable template DNA and has been successfully utilized in microsatellite markers (Barcellos et al., 1997) and SNPs (Shaw et al., 1998). Two bulks with 10 individuals each were generated from the mapping population. The first (R) and second (S) bulks encompass resistant and susceptible genotypes, respectively.

Genomic DNA was extracted from young leaf material with a commercial kits (Qiagen) according to the instructions of the manufacturer. DNA samples were quantified using the Qubit™ dsDNA BR Assay Kit with the Qubit® 2.0 Fluorometer (Invitrogen, Life Technologies, Darmstadt, Germany).

Genomic DNA from the parents of the mapping population V3125 x 'Börner', as well as from the two bulks, was used as template in 25 µl PCR reactions with the two marker-specific primers and 1 ng DNA per individual under standard conditions. The amplicons obtained were purified and sequenced from both directions using the ABI Prism BigDye Terminator chemistry on an ABI Prism 3730 sequencer (Applied Biosystems).

Determination of SNV/Allele Ratios

For the estimation of allele ratios, only bi-allelic BoeWGS1.0 SNVs that differentiate the two 'Börner' phases were considered, i.e., SNVs between BoeWGS1.0 contig pairs. Since only bi-allelic SNVs were considered, these positions are often homozygous in the V3125 genotype for one of the two 'Börner' alleles. Exceptions to these criteria (V3125 displays identical SNV heterozygosity to 'Börner') were excluded from the analysis. As a result, the allele frequency contributed from the 'Börner' parent has to be 50% and is 100% for the V3125 parent when counting the variant that is present in V3125. Following these assumptions, the expected frequency for each single allele in a bulk of the unselected F1 individuals of this mapping population is 75%. We have estimated the SNV frequencies within the parental genotypes in order to confirm the determination and to gain hints for problematic amplimers that amplify the haplophases with a bias for one of the two.

The estimation of the SNV frequencies among the pooled DNA samples (Pool R = resistant; Pool S = susceptible) and comparison with the parental lines was carried out by using the tool QVSanalyzer (Carr et al., 2009). The Sanger sequence trace files for determination of the relative proportions of two sequence variants were analyzed per batch. The generated output files contain details of the examined sequence variant ratios for individual samples as well as summary statistics. The area below the peak at the position of the targeted SNV was calculated and set into context with the surrounding peaks for each sample and the corresponding trace file. For graphical presentation of the results, the ratios were converted into percent values for SNV or allele frequency. SNVs considered in the analysis are listed in **Supplementary Table 3**).

RESULTS

A Draft Assembly of the 'Börner' Genome Sequence

The data for the 'Börner' genome sequence were obtained by whole genome shotgun (WGS) sequencing. The assembly, designated BoeWGS1.0, has a total size of 618.3 Mbp, and the N50 sequence size was 4,255 bp (**Supplementary Table 2**). We evaluated the average AD for the contigs in the assembly and classified the contigs with respect to the expected depth of aligned reads for paired contigs (separated haplophases) and contigs with merged sequences (both haplophases combined into one contig, see **Supplementary Figure 1**). A significant proportion of the contigs showed an AD between 50 and 100 (75 ± 25), which can be considered as the expected value for a haplophase-specific contig. The range has been deduced from the total amount of reads (75 Gbp), 2 x 500 Mbp for the expected fully diploid sequence length, and an interval of ±33%.

REAPR flagged a total of 14,007 regions (5.9%) as erroneous. Fragment coverage distribution errors within a contig accounted for 4,250 cases (1.8%), while the same error type over a gap added 5,674 cases (2.4%). The remaining cases were contributed by low coverage within a contig or across a gap with 3,723 (1.6%) and 360 (0.2%) cases, respectively. In summary, the assessment by REAPR indicated a short-read assembly from heterozygous material of acceptable to good quality. The heterozygosity of 'Börner' was calculated to be about 3.1% independent of the k-mer length used in the analysis. This is a quite high value in comparison to the *V. vinifera* variety 'Pinot noir' for which we calculated a heterozygosity of 1.5% regardless of k-mer length (**Supplementary Table 4**).

BUSCO, i.e., the Benchmarking of Universal Single-Copy Orthologs, revealed the presence of 55.9% of all 1,440 benchmarking genes from the reference set for embryophyta, 42% of the total 1,440 as single copy and 13.3% as duplicated. In addition, fragments of 19.1% benchmarking genes were detected. Only 25% of the benchmarking genes were not detected in the assembly. It should be noted that the duplicated BUSCO genes can be explained by detection of two allelic versions in BoeWGS1.0 contig pairs.

Taken together, the BoeWGS1.0 assembly represents a typical short-read assembly of a heterozygous genotype. Since we wanted to address marker development, we optimized the assembly for phase separation and not for continuity.

BoeWGS1.0 Contigs Representing Different Haplophases

The BoeWGS1.0 contigs were mapped against the *V. vinifera* reference genome sequence PN40024 12x.v2. About 210,444 (88.3%) could be mapped successfully with at least 30% of their length (https://doi.org/10.4119/unibi/2938185). The ratio of the total length of the mapped BoeWGS1.0 contigs to the length of the reference chromosomes is 1.14 on average (1.02 to 1.31; **Supplementary Table 5**), indicating uniform assembly quality and an overall homogeneous synteny between 'Börner' and PN40024. The median of the average contig ADs for all 19 chromosome mappings was 86.3 (**Supplementary Table 5**). Conspicuously, it is only the genetically unassigned chromosome "Ukn," or "unknown," where the mapped BoeWGS1.0 contigs have an AD of 248 on average. The remaining, unmapped contigs contain sequences that are too diverse from the PN40024 reference sequence to be mapped.

More than one quarter (29.2%; 142 Mbp) of the PN40024 reference sequence was covered by BoeWGS1.0 contig pairs, again suggesting that the two 'Börner' haplophases were partially separated in the assembly. Most of these contigs show on average

an AD in the expected range. Another quarter of the reference (28.7%; 139 Mbp) was covered by one BoeWGS1.0 contig, and these contigs often (but not always) display higher AD values. In these cases, either the two 'Börner' haplophases were merged into one contig during assembly, or only one of the two haplophases displayed sufficient similarity to become mapped.

Due to repetitive sequences, a small fraction of the PN40024 reference sequence (3.7%; 18 Mbp) was matched by more than two contigs. Nearly 38.5% (187 Mbp) of the reference sequence (including 3% N-stretches) was not covered by any stringently matching BoeWGS1.0 contig (**Figure 1**).

Homozygous and Heterozygous Variant Frequencies

Genomic variants including SNVs contribute not only to intraspecific diversity but they are also candidates for valuable molecular markers. By mapping 'Börner' NGS reads against the reference sequence and subsequent variant calling, almost 5 million highly reliable bi-allelic SNVs were detected (https://doi.org/10.4119/unibi/2938178 and https://doi.org/10.4119/unibi/2938180).

Half of the 4,996,490 SNVs are heterozygous SNVs (2,536,406), which means one of the BoeWGS1.0 contigs shows the same nucleotide as the reference at a given position. The other half (2,460,084) are homozygous SNVs, describing variants only existing between PN40024 and BoeWGS1.0. The frequency is on average 1 per 77 bp (**Supplementary Table 6**), ranging from 1/70 to 1/82 bp for chr.3, chr.8, and chr.16. In addition to SNVs, MNVs and small InDels were called, dropping the overall variant frequency slightly to 1 variant per 68 bp.

SSR Detection for Marker Development

Simple sequence repeat (SSR) markers are often used to study molecular diversity or heterozygosity as well as for genetic mapping in grapevine. By comparing SSR positions in the grapevine reference genome sequence with BoeWGS1.0, a total of 10,820 putative SSR marker positions ("candidates") with different unit sizes (two to six) and repeat numbers (up to 27) were deduced (**Table 1**). Regions of the reference matched by a BoeWGS1.0 contig pair were exploited. Out of the 10,820 positions, 12% (1,313) were monomorphic and 38% (4,110) were tri-allelic in an alignment of the three sequences. The remaining 50% were either bi-allelic or showed more than one motif in the SSR region. The more than 4,000 tri-allelic SSR candidates are the most valuable ones with regard to genetic mapping in a broad range of accessions (**Supplementary Table 7**). For various reasons, not all of them are new; for example, candidate SSR chr1_1203 corresponds to the established marker GF01-03, SSR chr1_1083 to marker GF01-21, and SSR chr1_1248 to marker GF01-22 (Fechter et al., 2014). However, the identification of already existing markers convincingly indicates that our candidate list is valid.

A set of 45 tri-allelic and eight bi-allelic SSR candidates with a unit size of two or three randomly selected from all

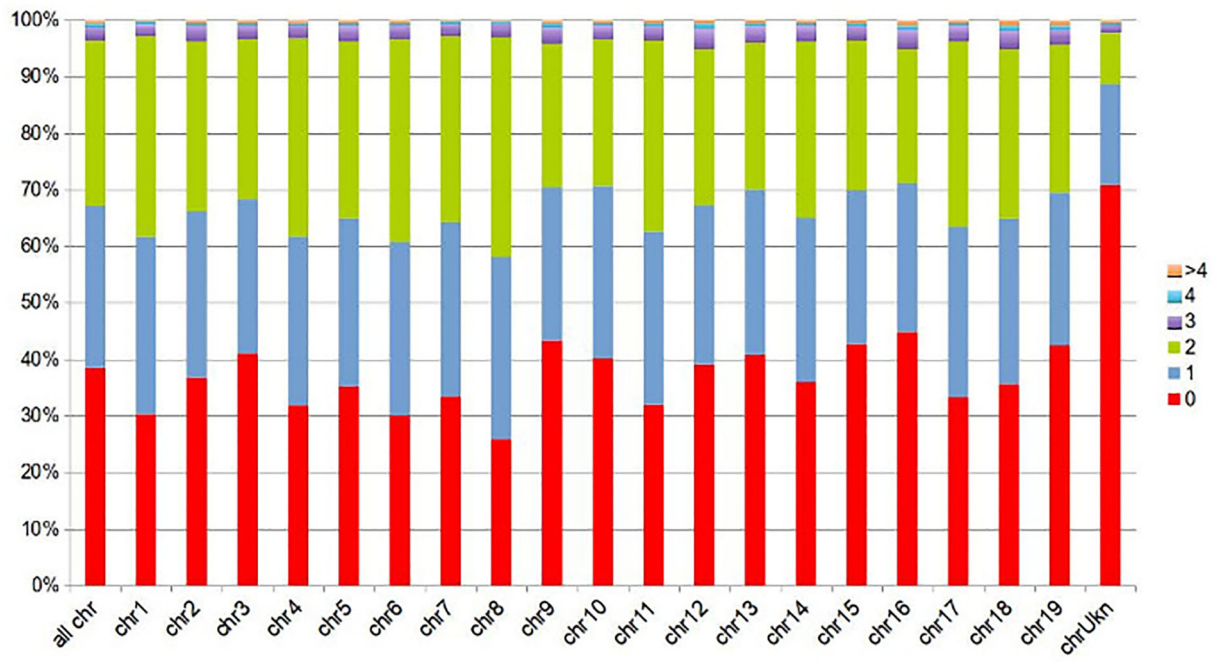

FIGURE 1 | Sequence fractions of the PN40024 reference covered by BoeWGS1.0 contigs, displayed for all PN40024 pseudochromosomes individually (chr#), and for the whole PN40024 reference (all chrs) in the leftmost column. Not covered fractions (given in percent) are shown in red (0), fractions covered by a single (1) BoeWGS1.0 contigs in blue, and fractions covered by contig pairs (2) in green. The remaining fractions are covered by three or more BoeWGS1.0 contigs (3, 4, > 4).

TABLE 1 | Mining and statistics of SSR containing sequences for marker development.

	PN40024_12Xv2	BoeWGS1.0
Total number of sequences examined *:	20	210444
Total size of examined sequences (bp):	486205130	539624286
Number of SSR containing sequences:	20	58397
Total number of identified SSRs:	**88635**	**95432**
Unit size number of SSRs = 2	54164	58573
Unit size number of SSRs = 3	26473	27723
Unit size number of SSRs = 4	6436	7477
Unit size number of SSRs = 5	1129	1197
Unit size number of SSRs = 6	433	462
Number of sequences containing more than 1 SSR:	20	21732
Number of SSRs present in compound formation**:	11861	9376
All matches of three SSR containing sequences with one being the reference		**10820**
Not polymorphic matches		1313
Bi-allelic matches		3226
Tri-allelic matches		**4110**
Tri-allelic matches with SSR consisting of >1 motif		2171
Total polymorphic matches		9507

*(excluding the fraction of BoeWGS1.0 that did not map to the PN40024 reference).
**(number of bases between two SSRs was ≤ 100).

chromosomes was used to validate the power and reliability of SSR candidate prediction. The bi-allelic candidates were included to estimate the chance of false negative predictions. In this subset of eight candidates, we tested six candidates; the reference matched only one contig of BoeWGS1.0, and two candidates were not found on a contig pair but shared a unit size and number. Out of the total set of 53 SSR sequences, 38 could be verified by either PCR and subsequent agarose gel-electrophoresis, Sanger sequencing, or fragment analysis (**Supplementary Table 8**). The 15 negative cases were caused by unspecific primers that amplified various fragments instead of allelic amplicons. We did not optimize the amplimers empirically in a second round of amplimer design.

As expected, 'Börner' amplicons with predicted small sequence variations between the alleles were analyzed more successful by Sanger sequencing than those with larger sequence variations. In six out of eight cases, the predicted bi-allelic SSR sequences were verified as monomorphic within 'Börner'. Analysis of the sequenced nucleotides before and after the SSR region confirmed the assumption that these genome regions were less heterozygous than others (or monomorphic), and they were therefore only represented by a single contig in the BoeWGS1.0 assembly. For the SSR chr1_43023 candidate, the new marker GF01-59 was established, although for this candidate only one BoeWGS1.0 contig matched to the reference. Both Sanger sequencing and fragment analysis confirmed the existence of two 'Börner' alleles for GF01-59.

Sanger sequencing of amplicons from SSR candidate loci with larger differences (e.g., in unit size, unit number, etc.) was less successful. In addition, we observed in these cases higher heterozygosity between the alleles even outside the SSR itself. However, validation by agarose gel-electrophoresis worked out smoothly for these SSR candidates.

Since gel-electrophoresis did not allow us to verify exact allele sizes, these were determined by fragment analysis in 14 cases (**Supplementary Table 8**). Fragment analysis was performed with DNA of the exact parental lines of 'Börner' and V3125 in addition to 'Börner' itself. For 12 out of 14 SSR candidates different fragment sizes for the two 'Börner' alleles were detected that fit well to the predictions for both 'Börner' alleles. The validated SSR candidates—for which we generated complete documentation—received marker designations. Seven of these SSR markers turned out to be highly or fully informative because they discriminate all four alleles of the biparental cross V3125 x 'Börner'. One marker (GF01-60, derived from candidate SSR chr1_100137_2) was scored as monomorphic in this analysis, likely due to the small sequence difference of 2 bp in fragment size. In fact, exactly the 2 bp difference in fragment length were properly detected by Sanger sequencing, these but are surely below the reliable resolution for a marker assay based on fragment analysis. The last of the 14 SSR markers, GF06-19 (SSR chr6_800_1), did not show any scoreable fragment for 'Börner,' but amplified a single scoreable amplicon in V3125. Taken together, every method seems to have its limitations, but the different validation approaches demonstrated the high reliability and usability of our prediction approach for new SSR markers. Further investigation of the highly informative markers in the grapevine varieties 'Schiava Grossa', 'Riesling', 'Villard Blanc', 'Calardis Musque' and 'Pinot noir', as well as in the parents of 'Börner' showed a transferability of about 70 to 80%. Most of the assayed SSR positions were heterozygous within and between the different cultivars.

Targeted Mapping of the Downy Mildew Resistance *Rpv14* by Bulk Segregant Analysis

To assist the introgression of the downy mildew locus *Rpv14*, which has been mapped in 'Börner' to the lower arm of chr.5 (Ochssner et al., 2016), into *V. vinifera*, closely linked markers are required. For SNV marker-based association studies for *Rpv14*, 25 amplimers were designed targeting the lower arm of chr.5 of 'Börner', and 17 of those turned out to be functional. As controls, two additional amplimers were designed, one at the top of chr.5 opposed to *Rpv14* and one on chr.14. All primers were designed based on BoeWGS1.0 contigs and SNV data (see Methods). The 76 amplicons obtained for 19 markers and 4 template DNAs (susceptible V3125, resistant donor 'Börner', and the F1 bulks S and R) were sequenced, and the trace files were subjected to peak area determination and evaluation.

All predicted SNVs within or between the BoeWGS1.0 contigs could be verified. In one case, we observed a tri-allelic SNV and removed the respective data points from further analyses. For SNVs from physically distinct contigs, the peak ratios or SNV frequencies varied as initially expected for a bulked segregant analysis. By ordering BoeWGS1.0 contig pairs along the chr.5 of the reference genome sequence and allocating the corresponding SNVs or allele frequencies to these contigs, the

changes in the frequencies of the pools are obviously correlating with the candidate locus (**Figure 2**). For the control regions on chr.14 and at the top of chr.5, we hypothesized no selective pressure and therefore expected an allele frequency for the bulks of 75%. The observed frequencies on chr.14 fulfilled the expectation. The frequencies in the north of chr.5 in the R bulk (55–60%) differ to some extent from the expectation where, as in the S bulk, the SNV frequency is as expected. This may be due to the low number of genotypes within the bulks. Even if only one out of the ten genotypes has no recombination between the target locus and the terminal part of the chromosome, the estimated frequency value changes theoretically by almost 10% of the expected value.

Within the region of the candidate locus between approximately 19 and 21 Mbp, the resistant bulk has a nearly uniform estimated SNV frequency of 50%. This indicates that all resistant individuals in the bulk carry the same allele of the resistant 'Börner' parent (inherited from *V. cinerea*) (Ochssner et al., 2016). Since there are only a limited number of recombination events possible, the large interval is not surprising. At the same time, the SNV frequency in the susceptible bulk converges (continuously) to about 95%. Our results indicate that the variants of 'Börner', located between 19.67 Mbp and 20.6 Mbp in the corresponding reference sequence PN40024 are highly and specifically linked to the resistant trait. The center of the candidate locus with the highest linkage, showing SNV frequencies over 85% for the S bulk is an interval of 330 kbp between 20.31 and 20.6 Mbp. This is the region flanked by the BoeWGS1.0 contigs c35489 and c244127 on one side, and c12059 and c12060 on the other (NCBI accessions CCJE01069526.1, CCJE01173868.1, CCJE01026433.1 and CCJE01045859.1).

DISCUSSION

Genome Sequencing

For the development of molecular markers and to gain knowledge about genome regions as well as loci relevant for grapevine breeding, we analyzed the genome of the resistant interspecific hybrid 'Börner' (*V. riparia* x *V. cinerea*) by WGS sequencing. The draft assembly BoeWGS1.0 covers a total of 618 Mbp, which corresponds to 62% of the expected 1 Gbp long diploid 'Börner' genome. Assessment by REAPR revealed some critical regions that could be caused by the incompletely resolved haplophases. Mapping of reads to at least in parts very similar sequences, such as separated alleles, still poses a challenge. However, most of the contigs pass the read mapping-based assembly evaluation.

A major challenge during the *de novo* assembly of this dataset was to optimize the separation of sequences into contigs derived from both haplophases for subsequent generation of highly informative SSR and SNV marker assays. Different evaluations for the success of this separation result in somehow related—but in detail different—numbers. BUSCO finds about 14% duplicated benchmarking genes. Some genes in the reference set are,

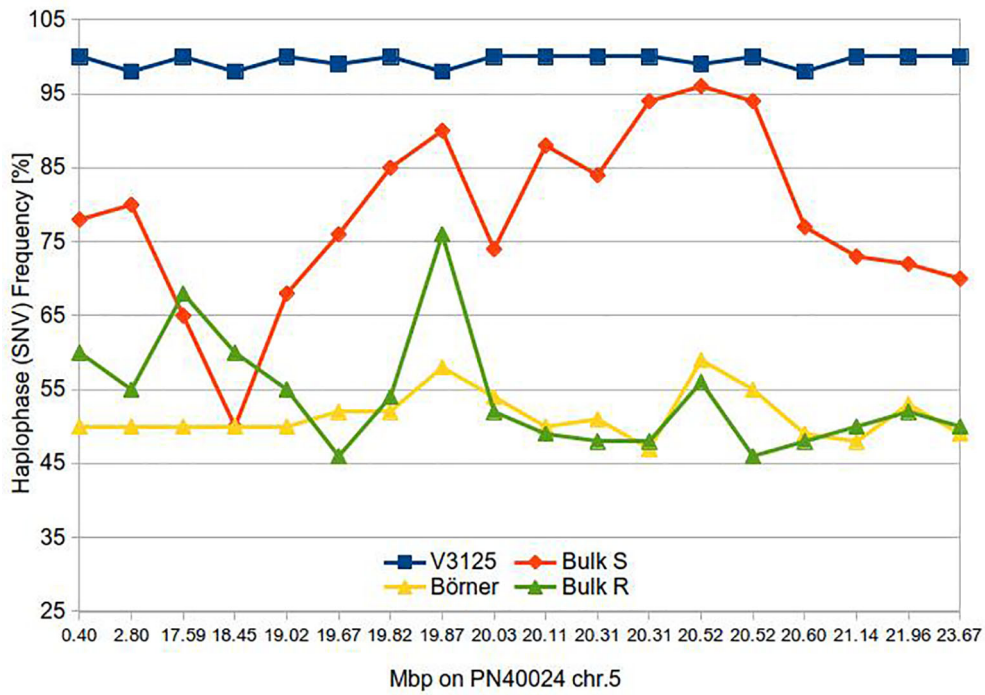

FIGURE 2 | Haplotype frequency along the *Rpv14* target region of parents and bulks of susceptible and resistance F1 individuals.

however, not detected at all (BUSCO completeness is about 75%). This is probably an underestimation for the contig level since the relatively small contigs complicate gene calling. The total length of the assembly is about 120 Mbp longer than the haploid *Vitis* genome sequence, indicating that 240 Mbp of the 618 Mbp represent sequences from separated haplophases. The analyses of contig mappings against the reference genome sequence PN40024 indicates that 142 Mbp of the reference are covered by two BoeWGS1.0 contigs. Taken together, we assumed that the BoeWGS1.0 assembly contains phase-separated sequence information for one quarter of the 'Börner' genome. Note that this information is not "phased throughout" since, even if a contig pair is detected, we lack information about which of the two belongs to the *V. cinerea* and which belongs to the *V. riparia* haplophase. Anyway, since the regions for which separation worked are distributed throughout the genome, the data are a very good source for marker development.

Although the separation of sequences from both haplophases is great for molecular breeding applications and several genome-wide investigations, the short-read-derived BoeWGS1.0 assembly is still a draft genome sequence. The short contig length significantly limits downstream analyses like gene prediction or approaches to detect genomic rearrangements. The main reason for the fragmentation of the BoeWGS1.0 assembly are regions where the ancestor genomes of *V. cinerea* and *V. riparia* are quite similar. In fact, the frequency of SNVs and MNVs detected after read mapping between the two haplophases of 'Börner' is in the same range as between 'Börner' and PN40024. However, it should be considered that read mapping requires quite similar sequences to be specific.

Currently, substantially longer read lengths (mean >20 kbp) than those that were used here are provided by third-generation single-molecule sequencing technologies like those offered by Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT). The long reads can be used to solve the problem of bridging repeat sequences or transposable elements. The first *de novo* assembly of domesticated grapevine using PacBio sequencing was from *V. vinifera* cv. 'Cabernet Sauvignon' (Chin et al., 2016) followed recently by the diploid assembly of the cultivar 'Chardonnay' (Roach et al., 2018; Zhou et al., 2019). The first and only wild *Vitis* genome sequence so far comes from the rootstock *Vitis riparia* 'Gloire de Montpellier' and reached considerable continuity by incorporating PacBio and 10X Chromium (Girollet et al., 2019). However, assembling a highly heterozygous genotype, accurate phasing, and aligning of *Vitis* haplotypes is still challenging. With regard to the data on hand and the complexity of this assembly, the application of dedicated scaffolding tools like scaffoldScaffolder (Bodily et al., 2015) or SSPACE (Boetzer et al., 2011) comes at a high risk of generating erroneous connections between contigs. Therefore, we calculated the BoeWGS1.0 assembly with very stringent parameters using the CLC Genomics Workbench toolkit, which contains an implementation of a de Bruijn Graph assembler, and refrained from scaffolding.

In addition to the room for improvement for the assembly parameters, there is a significant fraction of the BoeWGS1.0 genome sequence that is currently not able to align, and it is thus not assignable to the reference genome sequence. This genome fraction likely holds sequence information underlying important traits. Chromosomal anchoring of these contigs is a task that could be realized by the use of the provided SSR or SNV marker sequences of this study in an existing F1 mapping population (V3125 x 'Börner') (Fechter et al., 2014). Finally, the generation and incorporation of long read data for improvement of the 'Börner' genome sequence is an option for future studies. We expect that a long-read assembly with good continuity would also allow to study 'Börner' sequences that display high divergence from the reference.

Deduction of SSR Sequences for Marker Development

The detected polymorphic SSR candidate loci are a very valuable resource for genome-wide marker development and genotyping. SSR markers are still the most abundant molecular marker type used for mapping in *Vitis* species and marker-assisted selection in grapevine breeding programs. Because of their very polymorphic nature, SSRs often allow us to clearly distinguish between more than two alleles. This is highly necessary when using F1 mapping populations derived from a cross of highly heterozygous parents, when following an allele in a phylogenetic tree, or when identifying accessions (Cipriani et al., 2011). All these conditions fit to grapevine.

In this study, we were able to reliably predict thousands of candidates for SSR markers. Furthermore, we validated high prediction reliability, which is essentially limited only by the design of specific primers, and demonstrated their applicability in *Vitis* genotyping approaches. The prediction schema for selection of promising SSR candidates that relied on aligning contigs representing separated haplophases instead of large amounts of sequence reads turned out to be very successful. We were able to reliably detect, predict, and select for longer sequence variations caused by the variable number of repeated units. This is favorable for marker detection by fragment analysis as the current predominant method used for SSR-based genotyping.

Heterozygosity of 'Börner'

For *V. riparia* (cv. 'Gloire de Montpellier'), the mean distance between heterozygous SNVs is 217 bp (Girollet et al., 2019). For the heterozygous 'Pinot noir,' a frequency of 1 nucleotide variant per 100 bp and 1 InDel per 0.45 kbp was described (Velasco et al., 2007). The overall variant frequency that we obtained between the 'Börner' genome sequence and the reference PN40024 was 1 variant per 68 bp. This number is likely to be a significant underestimation because of the stringent read mapping parameters and the narrow confines for the coverage filter. A calculation of heterozygosity using GenomeScope indicated a heterozygosity of 3.1% for 'Börner' and 1.5% for 'Pinot noir' as an example for a prominent *V. vinifera* cultivar. This result gives additional indication for the highly heterozygous nature of the 'Börner' genome compared to *V. vinifera* varieties.

Application of Phased Contig Data for Targeted Mapping of the Downy Mildew Resistance *Rpv14* of *Vitis*

Single SNVs or groups of nearby SNVs within the multiple alignments of two 'Börner' contig pairs and the PN40024 reference sequence were shown to be useful for the molecular identification and localization of a quantitative trait locus. The highly reliable SNVs detected in our study can function as molecular markers in a bulked segregant analysis as shown for *Rpv14* on chr.5 in our analysis. *Rpv14* was selected as an example for proof of concept because the alternative *Rdv1*, which causes phylloxera resistance, has already been restricted to an area of about 350 kbp—a value deduced from the physical distance of the published flanking genetic markers (Zhang et al., 2009; Hausmann et al., 2014) relative to the PN40024 reference sequence. The other option, namely black rot resistance mediated by *Rgb1*/*Rgb2*, was not selected because there are at least two segregating loci (Rex et al., 2014). We were able to physically reduce the size of the *Rpv14* target region to less than 500 kbp. In relation to the results from Ochssner et al. (Ochssner et al., 2016), the center of the target region is narrowed down from the north and the south with the center being still at 20.06 Mbp. One limitation for further reduction of the target region was the number of F1 individuals in our R and S pools or BSA bulks. A recent review on BSA (Zou et al., 2016) mentioned an optimal bulk size for monogenic traits of 10 to 20% of the segregating population for each pool or "tail" of the trait value distribution. However, a BSA study in rice (Wambugu et al., 2018) was successful with 10 or fewer individuals per pool. In addition, equal size of the contrasting pools is considered to be important. In our study, the pool size of 10 resulted from the number of F1 genotypes with reliable and consistent phenotypic scoring results within the population. Obviously, this limits the resolution since only a relatively small number of recombination events were evaluated.

Further research should focus on the identification of new F1 individuals with additional recombinations in the target region. This will probably require enlargement of the number of F1 individuals in the mapping population and additional phenotyping for downy mildew resistance. Subsequently, the work presented here allows for quick access to additional markers, which can be used to determine the recombination points that then can be correlated with resistance genes from candidate gene predictions.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in https://www.ebi.ac.uk/ena/data/view/PRJEB28084, https://www.ebi.ac.uk/ena/data/view/PRJEB36326, https://doi.org/10.4119/unibi/2938185, https://doi.org/10.4119/unibi/2938178, https://doi.org/10.4119/unibi/2938180.

AUTHOR CONTRIBUTIONS

DH and LH conceived and planned the experiments. DH, PV, and LH designed and performed the experiments. TR, DH, and BP calculated the data. RT, BW, and DH conceived the original idea and supervised the project. DH, RT, and BW wrote the manuscript with input from all authors. All authors interpreted and discussed results.

FUNDING

The project was supported by the German Ministry of Education and Science (BMBF) grant FKZ 0315460A and 0315460B (Acronym: GrapeReSeq) to BW and RT, respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Willy Keller, Andreas Preiß, and Andreas Melnik for the excellent technical support. The authors also wish to thank the members of the Chair of Genetics and Genomics of Plants for their support. We gratefully acknowledge the financial support from BMBF through Projektträger Jülich. We also acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00156/full#supplementary-material

REFERENCES

Adam-Blondon, A.-F., Jaillon, O., Vezzulli, S., Zharkikh, A., Troggio, M., and Velasco, R. (2011). "Genome Sequence Initiatives," in *Genetics, Genomics and Breeding of Grapes*. Eds. A.-F. Adam-Blondon, J. M. Martinez-Zapater and C. Kole (Boca Raton: Science Publishers, CRC Press), 211–234. doi: 10.1201/b10948

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nat. Methods* 8 (1), 61–65. doi: 10.1038/nmeth.1527

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Barba, P., Cadle-Davidson, L., Harriman, J., Glaubitz, J. C., Brooks, S., Hyma, K., et al. (2014). Grapevine powdery mildew resistance and susceptibility loci identified on a high-resolution SNP map. *Theor. Appl. Genet.* 127 (1), 73–84. doi: 10.1007/s00122-013-2202-x

Barcellos, L. F., Klitz, W., Field, L. L., Tobias, R., Bowcock, A. M., Wilson, R., et al. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* 61 (3), 734–747. doi: 10.1086/515512

Bellin, D., Peressotti, E., Merdinoglu, D., Wiedemann-Merdinoglu, S., Adam-Blondon, A. F., Cipriani, G., et al. (2009). Resistance to Plasmopara viticola in grapevine 'Bianca' is controlled by a major dominant gene causing localised necrosis at the infection site. *Theor. Appl. Genet.* 120 (1), 163–176. doi: 10.1007/s00122-009-1167-2

Blasi, P., Blanc, S., Wiedemann-Merdinoglu, S., Prado, E., Ruhl, E. H., Mestre, P., et al. (2011). Construction of a reference linkage map of Vitis amurensis and genetic mapping of Rpv8, a locus conferring resistance to grapevine downy mildew. *Theor. Appl. Genet.* 123 (1), 43–53. doi: 10.1007/s00122-011-1565-0

Bodily, P. M., Fujimoto, M., Ortega, C., Okuda, N., Price, J. C., Clement, M. J., et al. (2015). Heterozygous genome assembly *via* binary classification of homologous sequence. *BMC Bioinf.* 16 Suppl 7. doi: 10.1186/1471-2105-16-S7-S5

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27 (4), 578–579. doi: 10.1093/bioinformatics/btq683

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31 (12), 1119–1125. doi: 10.1038/nbt.2727

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10. doi: 10.1186/1471-2105-10-421

Capistrano-Gossmann, G. G., Ries, D., Holtgräwe, D., Minoche, A., Kraft, T., Frerichmann, S. L. M., et al. (2017). Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. *Nat. Commun.* 8. doi: 10.1038/ncomms15708

Carr, I. M., Robinson, J. I., Dimitriou, R., Markham, A. F., Morgan, A. W., and Bonthron, D. T. (2009). Inferring relative proportions of DNA variants from sequencing electropherograms. *Bioinformatics* 25 (24), 3244–3250. doi: 10.1093/bioinformatics/btp583

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13 (12), 1050–1054. doi: 10.1038/nmeth.4035

Cipriani, G., Di Gaspero, G., Canaguier, A., Jusseaume, J., Tassin, J., Lemainque, A., et al. (2011). "Molecular Linkage Maps: Strategies, Resources and Achievements," in *Genetics, Genomics and Breeding of Grapes*. Eds. A.-F. Adam-Blondon, J. M. Martinez-Zapater and C. Kole. (Boca Raton: Science Publishers, CRC Press), 111–136. doi: 10.1201/b10948-6

Di Genova, A., Almeida, A. M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., et al. (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* 14. doi: 10.1186/1471-2229-14-7

Dohm, J. C., Lange, C., Holtgräwe, D., Sörensen, T. R., Borchardt, D., Schulz, B., et al. (2012). Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (Beta vulgaris). *Plant J.* 70 (3), 528–540. doi: 10.1111/j.1365-313X.2011.04898.x

Fechter, I., Hausmann, L., Zyprian, E., Daum, M., Holtgräwe, D., Weisshaar, B., et al. (2014). QTL analysis of flowering time and ripening traits suggests an impact of a genomic region on linkage group 1 in Vitis. *Theor. Appl. Genet.* 127 (9), 1857–1872. doi: 10.1007/s00122-014-2310-2

Girollet, N., Rubio, B., Lopez-Roques, C., Valiere, S., Ollat, N., and Bert, P. F. (2019). De novo phased assembly of the Vitis riparia grape genome. *Sci. Data* 6. doi: 10.1038/s41597-019-0133-3

Hausmann, L., Eibach, R., Zyprian, E., and Töpfer, R. (2014). Sequencing of the Phylloxera Resistance Locus Rdv1 of Cultivar 'Börner'. *Acta Hortic.* 1046, 73–78. doi: 10.17660/ActaHortic.2014.1046.7

Hausmann, L., Maul, E., Ganesch, A., and Töpfer, R. (2019). Overview of genetic loci for traits in grapevine and their integration into the VIVC database. *Acta Hortic.* 1248, 221–226. doi: 10.17660/ActaHortic.2019.1248.32

Herzog, E., Töpfer, R., Hausmann, L., Eibach, R., and Frisch, M. (2013). Selection strategies for marker-assisted background selection with chromosome-wise SSR multiplexes in pseudo-backcross programs for grapevine breeding. *Vitis* 52 (4), 193–196. doi: 10.5073/vitis.2013.52

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* 14 (5). doi: 10.1186/gb-2013-14-5-r47

Hyma, K. E., Barba, P., Wang, M., Londo, J. P., Acharya, C. B., Mitchell, S. E., et al. (2015). Heterozygous Mapping Strategy (HetMappS) for High Resolution Genotyping-By-Sequencing Markers: A Case Study in Grapevine. *PloS One* 10 (8). doi: 10.1371/journal.pone.0134880

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449 (7161), 463–467. doi: 10.1038/nature06148

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12 (4), 656–664. doi: 10.1101/gr.229202

Li, C., Lin, F., An, D., Wang, W., and Huang, R. (2018). Genome Sequencing and Assembly by Long Reads in Plants. *Genes* 9. doi: 10.3390/genes9010006

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.

Lodhi, M. A., and Reisch, B. I. (1995). Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. *Theor. Appl. Genet.* 90 (1), 11–16. doi: 10.1007/BF00220990

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27 (6), 764–770. doi: 10.1093/bioinformatics/btr011

Marguerit, E., Boury, C., Manicki, A., Donnart, M., Butterlin, G., Nemorin, A., et al. (2009). Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. *Theor. Appl. Genet.* 118 (7), 1261–1278. doi: 10.1007/s00122-009-0979-4

Merdinoglu, D., Wiedeman-Merdinoglu, S., Coste, P., Dumas, V., Haetty, S., Butterlin, G., et al. (2003). Genetic Analysis of Downy Mildew Resistance Derived from Muscadinia Rotundifolia. *Acta Hortic.* 603 (603), 451–456. doi: 10.17660/ActaHortic.2003.603.57

Michael, T. P., and Jackson, S. (2013). The First 50 Plant Genomes. *Plant Genome* 6 (2), 1–7. doi: 10.3835/plantgenome2013.03.0001in

Michelmore, R. W., Paran, I., and Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. U. States America* 88 (21), 9828–9832. doi: 10.1073/pnas.88.21.9828

Ochssner, I., Hausmann, L., and Töpfer, R. (2016). Rpv14, a new genetic source for Plasmopara viticola resistance conferred by Vitis cinerea. *Vitis* 55 (2), 79–81. doi: 10.5073/vitis.2016.55.79-81

Rex, F., Fechter, I., Hausmann, L., and Töpfer, R. (2014). QTL mapping of black rot (Guignardia bidwellii) resistance in the grapevine rootstock 'Börner' (V. riparia Gm183 × V. cinerea Arnold). *Theor. Appl. Genet.* 127 (7), 1667–1677. doi: 10.1007/s00122-014-2329-4

Roach, M. J., Johnson, D. L., Bohlmann, J., van Vuuren, H. J. J., Jones, S. J. M., Pretorius, I. S., et al. (2018). Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PloS Genet.* 14 (11). doi: 10.1371/journal.pgen.1007807

Schwander, F., Eibach, R., Fechter, I., Hausmann, L., Zyprian, E., and Töpfer, R. (2012). Rpv10: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. *Theor. Appl. Genet.* 124 (1), 163–176. doi: 10.1007/s00122-011-1695-4

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* 3 (11), 862–871. doi: 10.1038/nrg930

Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G., and Chakravarti, A. (1998). Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* 8 (2), 111–123. doi: 10.1101/gr.8.2.111

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31 (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Töpfer, R., and Eibach, R. (2016). Pests and diseases: Breeding the next-generation disease-resistant grapevine varieties. *Wine Vitic. J.* 31 (5), 47–49.

Taheri, S., Lee Abdullah, T., Yusop, M. R., Hanafi, M. M., Sahebi, M., Azizi, P., et al. (2018). Mining and Development of Novel SSR Markers Using Next Generation Sequencing (NGS) Data in Plants. *Molecules* 23 (2). doi: 10.3390/molecules23020399

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-

markers in barley (Hordeum vulgare L.). *Theor. Appl. Genet.* 106 (3), 411–422. doi: 10.1007/s00122-002-1031-0

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 (15). doi: 10.1093/nar/gks596

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature* 527 (7579), 508–511. doi: 10.1038/nature15714

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS One* 2 (12). doi: 10.1371/journal.pone.0001326

Venuti, S., Copetti, D., Foria, S., Falginella, L., Hoffmann, S., Bellin, D., et al. (2013). Historical introgression of the downy mildew resistance gene Rpv12 from the Asian species Vitis amurensis into grapevine varieties. *PloS One* 8 (4). doi: 10.1371/journal.pone.0061228

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33 (14), 2202–2204. doi: 10.1093/bioinformatics/btx153

Wambugu, P., Ndjiondjop, M. N., Furtado, A., and Henry, R. (2018). Sequencing of bulks of segregants allows dissection of genetic control of amylose content in rice. *Plant Biotechnol. J.* 16 (1), 100–110. doi: 10.1111/pbi.12752

Welter, L. J., Göktürk-Baydar, N., Akkurt, M., Maul, E., Eibach, R., Töpfer, R., et al. (2007). Genetic mapping and localization of quantitative trait loci affecting fungal disease resistance and leaf morphology in grapevine (Vitis vinifera L). *Mol. Breed.* 20, 359–374. doi: 10.1007/s11032-007-9097-7

Zhang, J., Hausmann, L., Eibach, R., Welter, L. J., Töpfer, R., and Zyprian, E. M. (2009). A framework map from grapevine V3125 (Vitis vinifera 'Schiava grossa' x 'Riesling') x rootstock cultivar 'Börner' (Vitis riparia x Vitis cinerea) to localize genetic determinants of phylloxera root resistance. *Theor. Appl. Genet.* 119 (6), 1039–1051. doi: 10.1007/s00122-009-1107-1

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., et al. (2019). The population genetics of structural variants in grapevine domestication. *Nat. Plants* 5 (9), 965–979. doi: 10.1038/s41477-019-0507-8

Zimmer, R., and Verrinder Gibbins, A. M. (1997). Construction and characterization of a large-fragment chicken bacterial artificial chromosome library. *Genomics* 42 (2), 217–226. doi: 10.1006/geno.1997.4738

Zou, C., Wang, P., and Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol. J.* 14 (10), 1941–1955. doi: 10.1111/pbi.12559

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 Plant Science

ORIGINAL RESEARCH
published: 25 March 2020
doi: 10.3389/fpls.2020.00287

Analyzing the Dietary Diary of Bumble Bee

Robert M. Leidenfrost[1], Svenja Bänsch[2]*, Lisa Prudnikow[1], Bertram Brenig[3], Catrin Westphal[2] and Röbbe Wünschiers[1]*

[1] Department of Biotechnology and Chemistry, Mittweida University of Applied Sciences, Mittweida, Germany, [2] Functional Agrobiodiversity, Department of Crop Sciences, University of Göttingen, Göttingen, Germany, [3] Institute of Veterinary Medicine, University of Göttingen, Göttingen, Germany

Bumble bees are important crop pollinators and provide important pollination services to their respective ecosystems. Their pollen diet and thus food preferences can be characterized through nucleic acid sequence analysis. We present ITS2 amplicon sequence data from pollen collected by bumble bees. The pollen was collected from six different bumble bee colonies that were placed in independent agricultural landscapes. We compared next-generation (Illumina), i.e., short-read, and third-generation (Nanopore), i.e., MinION, sequencing techniques. MinION data were preprocessed using traditional and Nanopore specific tools for comparative analysis and were evaluated in comparison to short-read sequence data with conventional processing. Based on the results, the dietary diary of bumble bee in the studied landscapes can be identified. It is known that short reads generated by next-generation sequencers have the advantage of higher quality scores while Nanopore yields longer read lengths. We show that assignments to taxonomic units yield comparable results when querying against an ITS2-specific sequence database. Thus, lower sequence quality is compensated by longer read lengths. However, the Nanopore technology is improving in terms of data quality, much cheaper, and suitable for portable applications. With respect to the studied agricultural landscapes we found that bumble bees require higher plant diversity than only crops to fulfill their foraging requirements.

Keywords: biodiversity, ecology, pollen, bumble bee, ITS2, next-generation sequencing, third-generation sequencing, Nanopore

OPEN ACCESS

Edited by:
Uwe Scholz,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
Jonathan Berenguer Koch,
University of Hawai'i at Hilo, United States
Mario Sergio Palma,
Paulista State University, Brazil

****Correspondence:***
Robert M. Leidenfrost
robert.leidenfrost@hs-mittweida.de
Svenja Bänsch
svenja.baensch@agr.uni-goettingen.de

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *25 November 2019*
Accepted: *26 February 2020*
Published: *25 March 2020*

Citation:
Leidenfrost RM, Bänsch S, Prudnikow L, Brenig B, Westphal C and Wünschiers R (2020) Analyzing the Dietary Diary of Bumble Bee. Front. Plant Sci. 11:287. doi: 10.3389/fpls.2020.00287

INTRODUCTION

Crop pollinators such as wild and domestic bees are important ecosystem service providers and in high demand (Aizen et al., 2008). The pollination services rendered by these pollinators are affected by changes to floral resources in semi-natural habitats and simplification of agricultural landscapes (Steffan-Dewenter and Westphal, 2008). Intensification of agricultural land use at local and landscape scales is considered as one major driver of pollinator declines due to shortages in the supply with pollen and nectar resources (Potts et al., 2010; IPBES, 2019). To sustain future crop pollination services in changing agricultural landscapes, it is important to characterize the foraging ecology of wild and domestic bees. Bumble bees are important crop pollinators because of their general floral diets and their large foraging ranges (Westphal et al., 2006; Kleijn et al., 2015).

We aim to identify the pollen diet of a common bumble bee species (*Bombus terrestris* L.) in agricultural landscapes. In this respect, the identification of pollen resources can reveal part of their food plant preferences and dietary requirements and thus can guide future conservation measures

and EU agri-environmental schemes. Identification and quantification is generally possible by labor-intensive pollen microscopy (Marzinzig et al., 2018) or nucleic acid sequence analysis (Danner et al., 2016). Approaches to the latter include DNA barcoding (Taberlet et al., 2012; Sickel et al., 2015; Bell et al., 2016) and genome skimming (Dodsworth, 2015). Most recently, a semi-quantitative approach involving Nanopore sequencing has been reported (Peel et al., 2019). The internal transcribed spacer (ITS) sequence is a popular genetic species barcode in plants (Chen et al., 2010; Yao et al., 2010; Bell et al., 2016).

In this study, we are sequencing ITS2-derived amplicons from plant pollen collected by bumble bees in order to identify pollen source species. From this data we derive bumble bees' pollen foraging under given environmental settings using a geographically customized BLAST database derived from the ITS2 database (Merget et al., 2012). Since ITS2-amplicons generated with common primer pairs typically exceed the length of polymerase-derived NGS-reads, we are evaluating full-length MinION-based ITS2-amplicon sequencing in contrast to NextSeq-based sequencing. From a technical perspective this work aims at developing field protocols for a rapid MinION-based assessment of pollen plant diversity in the field and utilization by pollinators, including estimation of crop pollination services delivered (Pomerantz et al., 2018; Krehenwinkel et al., 2019).

MATERIALS AND METHODS

Pollen-DNA extracts were PCR-amplified with ITS2-specific primers. Amplicons were then sequenced on NextSeq and MinION platforms, respectively (**Figure 1**).

Pollen Collection

Pollen was collected from bumble bees in front of their hives between May and June 2017. The bumble bee colonies were purchased from a German bumble bee breeder (STB Control, Aarbergen, Germany). The hives were located close to commercial strawberry fields (**Supplementary Material S1**). Pollen loads were collected by capturing, if possible, five individual bees in front of their colonies with an insect net. Pollen was removed from the hind tibia with tweezers. Afterward, bumble bees were released. We pooled the pollen loads of each observation date by colony and homogenized them in 70% (v/v) ethanol [one part pollen and four parts 70% (v/v) ethanol]. We prepared 1 mL aliquots for microscopic (not shown) and molecular pollen analysis by centrifugation for 10 min at 15,400 × *g*. We then removed the supernatant and dried them for 72 h in a clean bench.

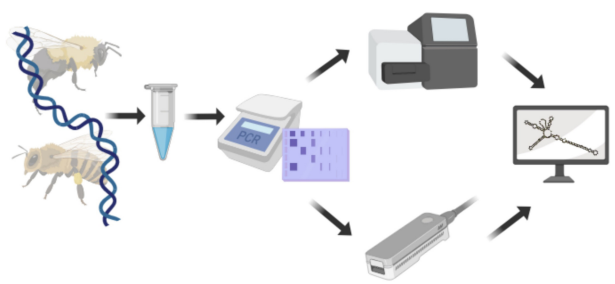

FIGURE 1 | Experimental setup used to compare Illumina and Nanopore sequencing technologies. DNA was extracted from pollen and the ITS2 region amplified. Amplicons were sequenced with either, Illumina NextSeq or Nanopore MinION sequencer before being subjected to analysis. Created with biorender.com.

Nucleic Acid Extraction

The DNA of approximately 0.015 g pollen aliquots was isolated using the DNeasy Plant Mini Extraction Kit from Qiagen according to the manufacturer's instructions. Cell lysis and homogenization of the samples were modified as follows: 150 g ceramic beads (1.4 mm), one tungsten carbide bead (3 mm), and 200 µL buffer AP1 were added to each dried sample. Samples were homogenized twice with a FastPrep Instrument (FastPrep® FP120, ThermoSavant) for 45 s at 6.5 m/s with a cooling step with ice in-between. Another 200 µL buffer AP1 were added. Finally, the standard protocol was followed until the DNA was eluted with 50 µL of elution buffer. DNA concentration and quality were measured using a Nanodrop (Thermo Fisher Scientific, Massachusetts, United States), and, prior to MinION Nanopore sequencing, with Qubit 3.0, dsDNA HS Assay Kit (Invitrogen, Eugene, United States).

ITS2 Amplicon Generation

For each sample, we performed three separate 10 µL PCR reactions to reduce PCR bias (Sickel et al., 2015) using the primers ITS2F [ATGCGATACTTGGTGTGAAT; Tm 61°C (Chen et al., 2010)] and ITS4R [TCCTCCGCTTATTGATATGC; Tm 60°C (White et al., 1990)]. Each reaction contained 0.3 µL FastStartTaq Polymerase (5 U/µL, Roche, Mannheim, Germany), 0.5 µL dNTPs (0.5 mM), 0.75 µL of each forward and reverse primer (10 pmol/µL), 2.5 µL 10× PCR buffer with $MgCl_2$ at a concentration of 20 mM (Roche, Mannheim, Germany), 19.2 µL PCR grade water, and 1 µL DNA template. The PCR conditions were optimized to the following conditions: initial denaturation at 95°C for 10 min, 37 cycles of denaturation at 95°C for 40 s, annealing at 49°C for 40 s, and elongation at 72°C for 40 s. Final extension was performed at 72°C for 5 min.

All reactions were checked for successful amplifications and contaminations by gel electrophoresis (1.5% agarose gels stained with ethidium bromide, 120 V for 30 min). Triplicate PCR products were pooled per sample and purified using the QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany).

NextSeq500 Illumina Sequencing

Paired-end sequencing (2 × 150 bp) was performed on a NextSeq500 platform (Illumina, San Diego, CA, United States) using a Mid-output flowcell (150 cycles). Of each amplicon 500 ng was used for library preparation according to the manufacturer's protocol (NEBNext Ultra II DNA Library Prep Kit for Illumina, New England Biolabs, Munich, Germany).

MinION Nanopore Sequencing

Nanopore sequencing of each amplicon was performed using the MinION [Oxford Nanopore Technologies (ONT), Oxford, United Kingdom] and 1D native barcoding according to protocols (EXP-NBD103 and SQK-LSK108, ONT; NEBNext End repair/dA-tailing Module, NEB Blunt/TA Ligase Master Mix, NEBNext Quick Ligation Module, New England Biolabs, Munich, Germany; AMPure XP beads, Agencourt) on a R9.4.1 flow cell (FAH89141, ONT, run QC = 1253 pores). Shearing and DNA repair steps were omitted. Incubation times during end-prep step were prolonged to 20 min. At designated checkpoints during library preparation, DNA was quantified using Qubit 3.0 fluorometer (dsDNA HS Assay Kit, Invitrogen, Eugene, United States). Data acquisition was performed by MinKNOW (v_1.15.6, ONT) and subsequent base-calling by Albacore (v_2.3.4, ONT).

Data Analysis

Basecalled MinION data were demultiplexed using Porechop (v_0.2.4, no further parameters set[1]) and assessed by NanoPack [v_1.13.0; Nanoplot 1.27.0 (Coster et al., 2018)]. A cursory look into the data was performed using Kraken2 [v_2.0.7-beta; NCBI non-redundant nucleotide database built in 2018-09 (Wood et al., 2019)] and subsequent visualization with Krona (Ondov et al., 2011). Reads were further processed by removing primers, using either USEARCH [v_11.0.667_i86linux32 (Edgar, 2010)] or Porechop containing ITS2F and ITS4R primer sequences.

In order to increase the accuracy of assignment of amplicon reads to plant-specific ITS2 sequences, we extracted all ITS2 sequences from a global eukaryota database (Ankenbrand et al., 2015) for plants that have previously been detected in Lower Saxony, Germany (Garve, 2004, 2007). The resulting subset was made non-redundant by clustering identical entries with VSEARCH (Version 2.9.1; Rognes et al., 2016) and subsequently used to create a magicBLAST database (version 1.4; Boratyn et al., 2018). After querying the Illumina amplicon reads against this database, all paired reads that both aligned to the same plant ITS2 sequence database entry with at least 50 bp each and a similarity greater than 98% were kept.

For each matching read, we calculated an alignment quality score by multiplying the alignment length with the alignment identity, thus accounting for overall alignment quality. The scores

[1]https://github.com/rrwick/Porechop

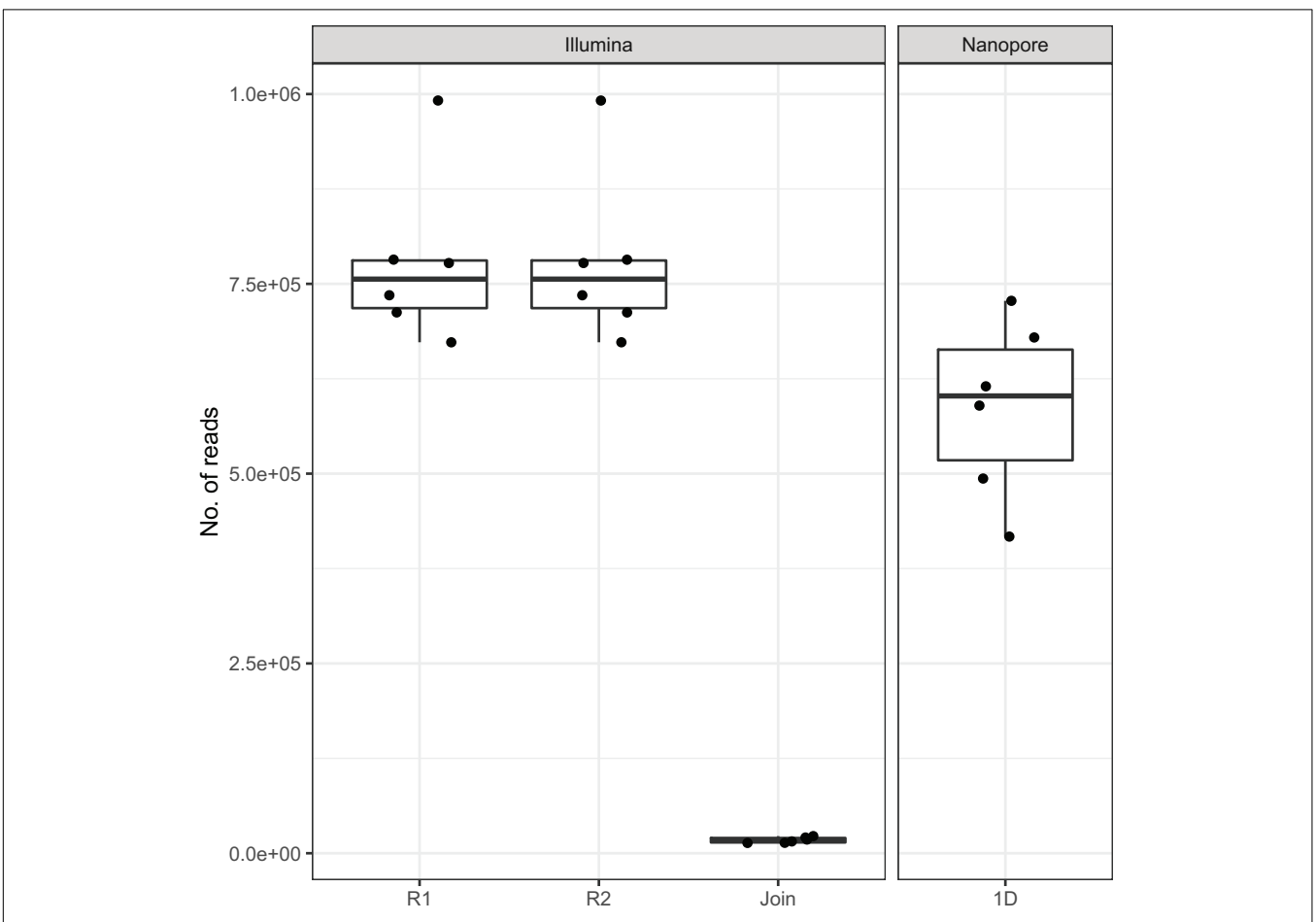

FIGURE 2 | Amount of reads generated with Illumina and Nanopore sequencers across all six samples. Note the low amount of joinable sequences for the Illumina data as a result of amplicon length > 300 bp.

for the forward and reverse read were summed to get a final score for each read-pair. Read-pairs that matched several entries were ordered by this score. Only the top scoring match (plant species) per read was counted. As some plant species have very similar ITS2 sequences and we, therefore, cannot unambiguously distinguish them on a species level, we decided to use all sequence data down to the genus level only. If there were more than one scoring match with an identical score, we decided on a match with higher reliability based on personal observations in the field, flowering time and a distribution atlas of plants in Lower Saxony (Garve, 2007). The final alignment quality score assigned to each read, respectively, was used for taxonomic assignment. Ultimately, pollen richness was calculated as the amount of plant genera in the respective pollen sample.

RESULTS AND DISCUSSION

On average, we retrieved 778,566 reads from the NextSeq and 588,252 reads from the MinION platform, respectively (**Figure 2**). While the read length was fixed to 150 nt by the Illumina chemistry, Nanopore reads varied from 340 to 380 nt with an average of 354 nt, after trimming with Porechop (**Figure 3A**). Generally, trimming reduced the average length of a MinION read by 25%, while at the same time increasing the average read quality score by 3.5%. A full length native barcode adapter, as identified by Porechop, is of ∼65 nt length, with the actual barcode consisting of 24 nt. Our trimming approach (using default parameters) resulted in the least removal of problematic artifacts and was made intentionally to establish a baseline. It may of course be made more stringent through more careful filtering before proceeding with downstream analysis, solving potential inaccuracies and circumventing technical artifacts (White et al., 2017; Xu et al., 2018).

With Nanopore technology being capable to generate sequence reads of several thousand nucleotides, the resulting average of 354 nt resembles full amplicon length (**Figure 3A**). Hence, it can be concluded that Illumina, even with paired-end sequencing, would not cover the whole amplicon, whereof 2 × 40 nt account for the forward and reverse primer, respectively. Even the sequencing kit for 300 cycles, at almost twice the cost, would be insufficient to provide full-length amplicon reads. Plant ITS2-sequences may exceed 600 nt (Yao et al., 2010). Therefore, only 3%, i.e., in average 17,406 of the paired-end reads, could be joined with standard bioinformatic tools [FastQ-join (Aronesty, 2013)] to full amplicon reads (**Figure 2**). Hence, we developed a magicBLAST pipeline as described in the methods to assign unjoined reads to target plants.

We observed only a fraction (5.9%) of Illumina reads that were shorter than the expected 150 nt. In contrast, Nanopore reads had a wider length variability (**Figure 3A**, min: 5bp; median: 350 bp; max. 11,519 bp), probably reflecting (a) varying ITS2-sequence sizes (Yao et al., 2010), (b) incomplete and/or unspecific amplicons, and (c) library preparation artifact. The latter is most likely based on the library preparation ligation protocol, since randomly picked long reads turned out to be concatenated amplicons.

The amplicon read mean quality scores (Phred score) were averaging around 30 for NextSeq data, which is approximately 15 to 20 units higher compared to the MinION data (**Figure 3B**). While the quality of reads generated by Nanopore sequencer technology can be expected to improve due to technical

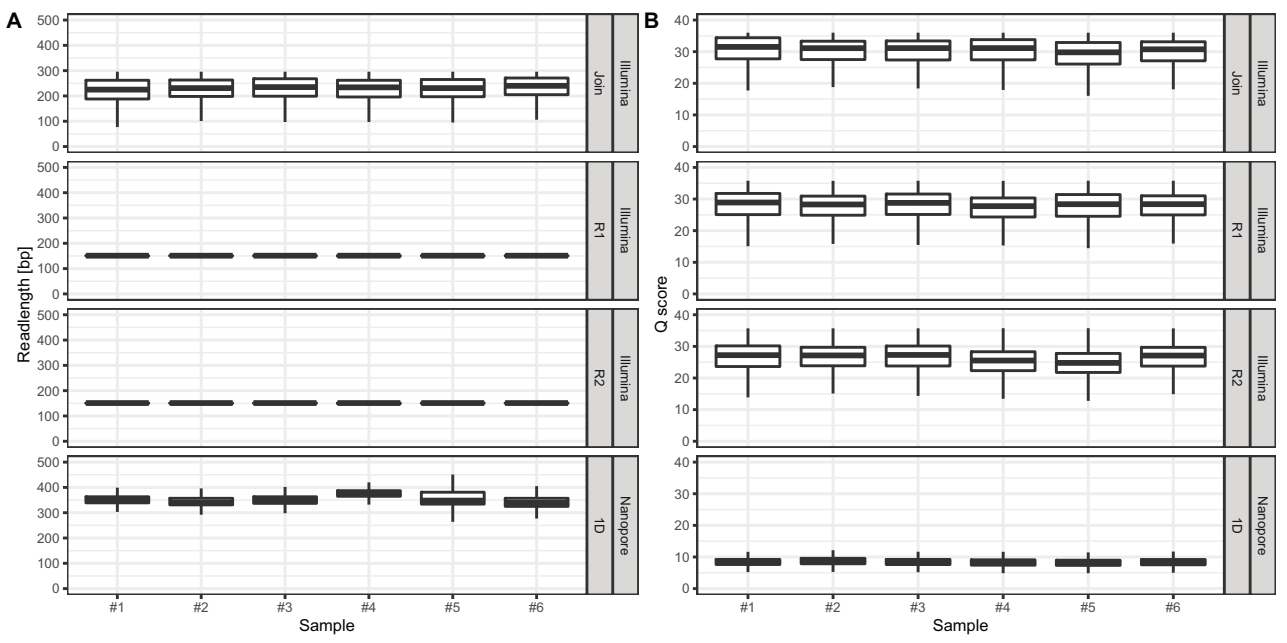

FIGURE 3 | (A) Read lengths [bp] for Illumina (min: 25, max: 296, figure split by R1, R2, and joined reads, respectively) and Nanopore (min: 5, max: 11,519) data per sample. **(B)** Corresponding mean read Q scores for Illumina (median: 27.5) and Nanopore (median: 8.5) data per sample. Outliers are not displayed.

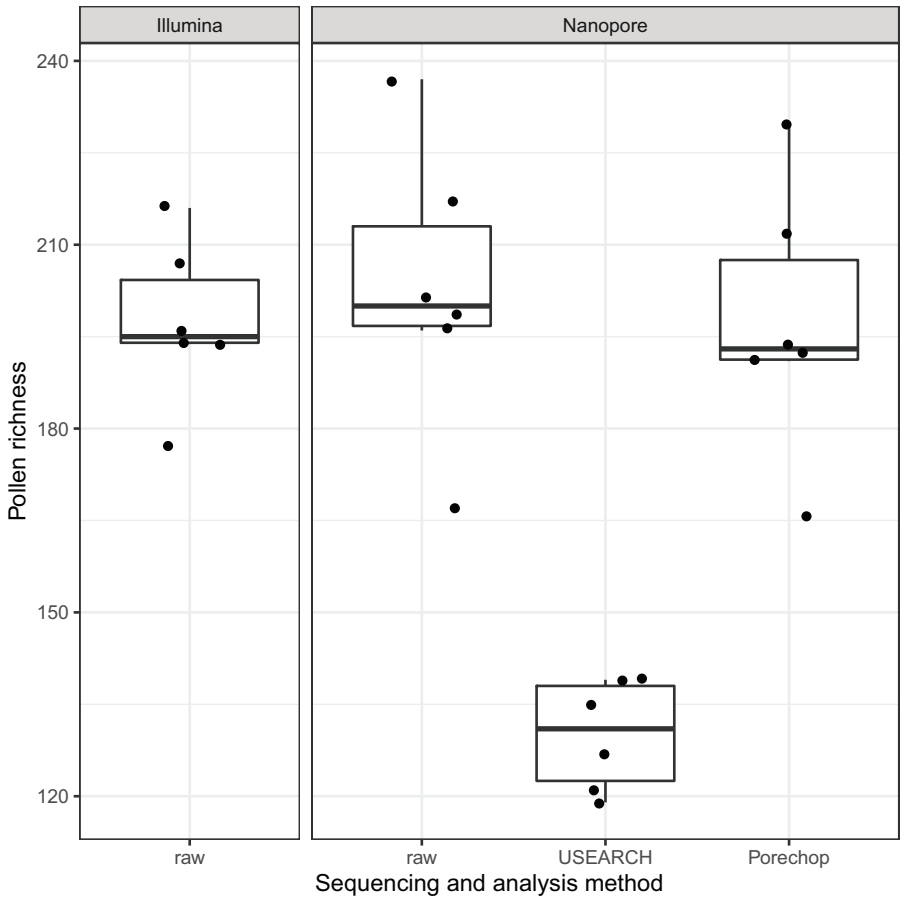

FIGURE 4 | Boxplot-comparison of assigned genera for Illumina and Nanopore reads, respectively. Pollen richness corresponds to the number of genera detected when BLAST querying against the Lower Saxony specific ITS2-sequence database. USEARCH's amplicon extraction led to less assignments than Porechop's read-end trimming.

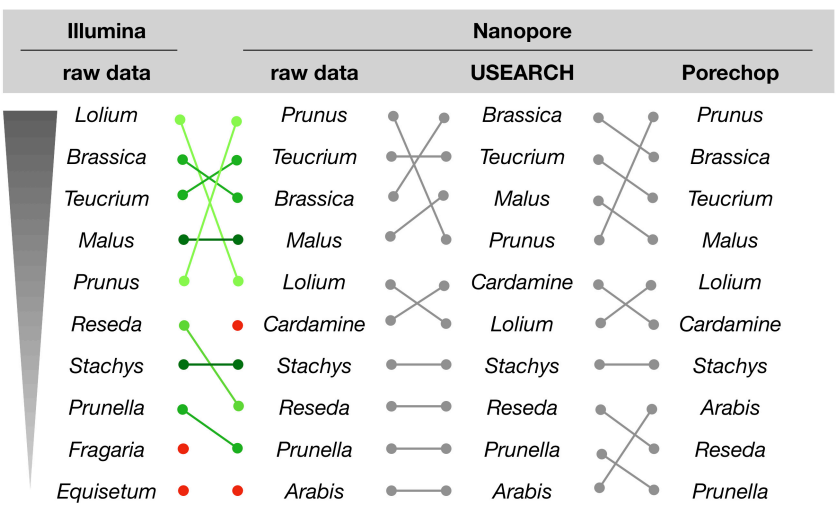

FIGURE 5 | Genus assignment for sample #6 by querying raw or processed sequence reads against the Lower Saxony specific ITS2-sequence database. Only the ten most abundant genera are shown in descending order, as indicated. Lines represent different ordering; red dots indicate genera that appear further down in the list.

optimizations, at the current technical level, short reads generated by next-generation sequencer technology such as provided by Illumina are of better quality (Rang et al., 2018; van Dijk et al., 2018). The lower average read quality scores of the sequence reads generated by Nanopore MinION reflect its error prone nature. This is especially the case with the flanking regions containing the 20 nt primer sequences. Those contained up to 30% single nucleotide mismatches. Yet, Nanopore reads can still be BLAST-assigned to the ITS2-sequence database to a similar extent as Illumina reads: Average pollen richness, i. e. assigned genera, for Illumina reads is 197 (min.: 177; max.: 216). For Nanopore reads the average pollen richness is 203 (min.: 167; max.: 237) (**Figure 4**). Primer clipping with Porechop does hardly change the mean pollen richness, albeit a wider span is observable (mean: 198; min.: 166; max.: 230). In contrast, amplicon extraction with USEARCH reduces the number of assignments (mean: 130; min.: 119; and max.: 139).

For the initial comparison of sequencing technologies as presented in this study, we focused on a qualitative rather than quantitative analysis of the assignment results. With respect to the genus assignments performed on NextSeq and MinION data, the sample with the most divergent ranking (sample #6), is differing only in the order, but not the presence of the ten most abundant genera (**Figure 5**). This result is supported by a microscopic analysis of pollen grains (not shown).

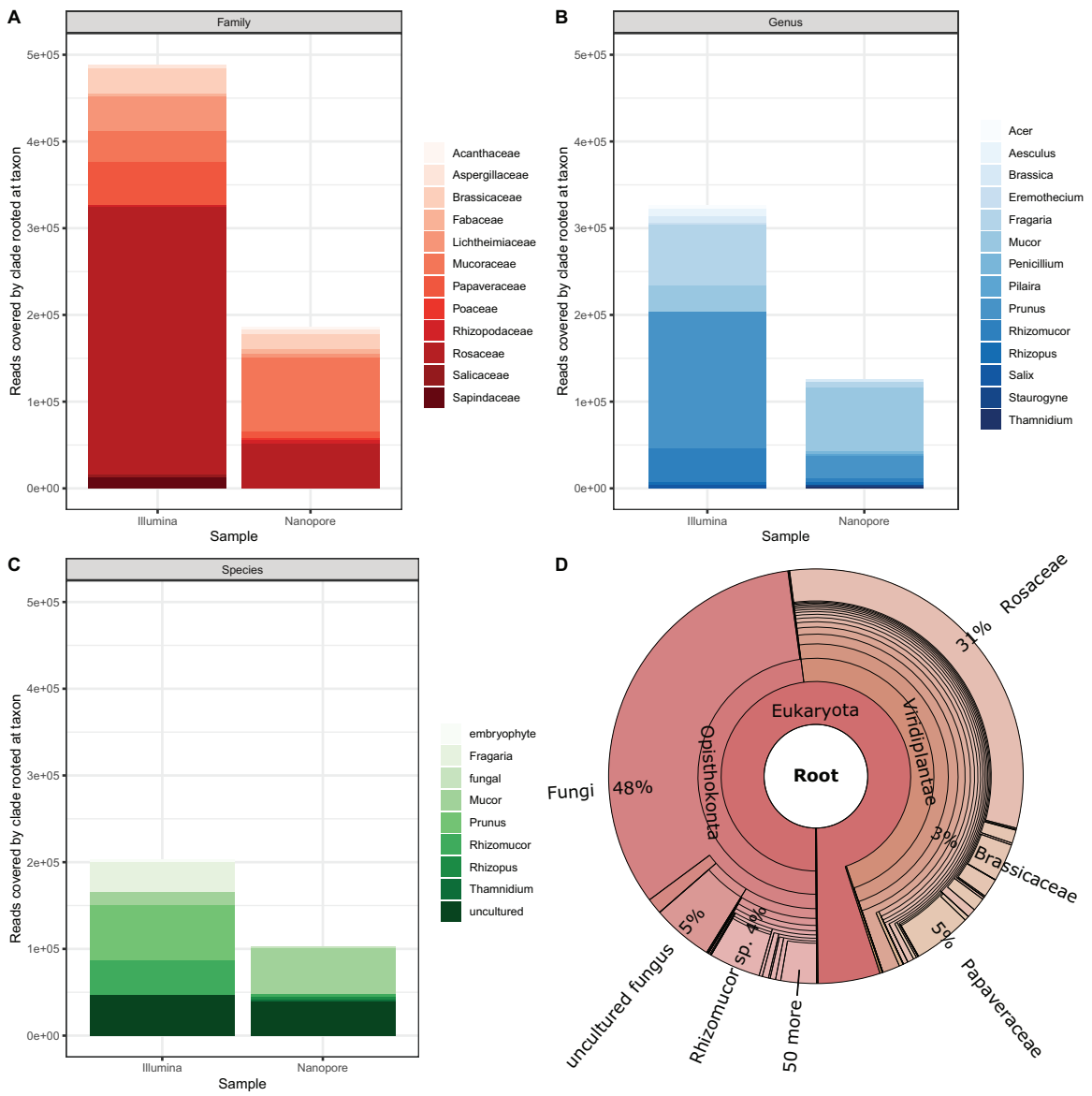

FIGURE 6 | Cursory display of the ten most abundant assignments from Kraken2 reports for sample #6 data generated by Illumina and Nanopore platforms, respectively on **(A)** Family, **(B)** Genus, and **(C)** Species level. **(D)** Krona plot for the Illumina data (17 levels displayed).

Finally, we applied an assignment approach without the application of BLAST and our Lower Saxony specific ITS2-sequence database. Instead we used Kraken2 that queries for exact k-mer sequence matches in a k-mer database that is based on NCBIs non-redundant nucleotide DB. This approach achieves high accuracy with fast classification speed (Wood and Salzberg, 2014; Wood et al., 2019). Again, to establish a baseline, we focus on the sample that generated most divergent results between Illumina and Nanopore data (sample #6, **Figure 6**). Prominently, taxonomic units other than plants are detected as a result of the DB employed by Kraken2. This "bycatch" constitutes representatives from the kingdoms fungi and – in lower abundance – metazoa and bacteria. While caution must be taken when interpreting this finding in detail for the Nanopore data due to their error-prone nature, the detection of especially fungal species was also clearly visible in the Illumina data visualized with Krona (**Figure 6D**). Indeed, the presence of molds in pollen is not uncommon (Kacániová et al., 2009; Belhadj et al., 2015; Nardoni et al., 2016). Moreover, despite Nanopore yielding less than half of the total reads (Illumina ∼990k reads, Nanopore ∼417k reads), Kraken2 assigned those reads to roughly twice the assigned genera (Illumina ∼3,648 genera, Nanopore ∼1,731 genera). This is, again, most likely due to (a) the error prone nature of Nanopore reads (Rang et al., 2018), and (b) the much larger database size (NCBI non-redundant nucleotide k-mer database versus Lower Saxony plant specific ITS2-sequences). Ultimately, the choice for either approach, ITS2 versus Kraken2, depends on the research purpose.

In terms of bumble bee foraging in different agricultural landscapes, our results show that colonies are not only heading to the close strawberry field (*Fragaria*). Instead, also plants of the genus *Brassica*, which is most likely oilseed rape because it is flowering intensively in May in the investigated regions, and flowers of a great variety of other plant genera were visited (**Figure 6B**). Beside the annual crops (e.g., oilseed rape and strawberry) in the agricultural landscape matrix, bumble bees also visited woody structures such as *Prunus* and *Acer*. Cherry trees belong to the genus *Prunus* and are commonly found in home gardens but also along roadsides. The same is true for *Acer*, *Aesculus* (chestnut), and *Salix* (willow), which are common trees in agricultural and urban areas. Our findings indicate that bumble bees visit much more plants genera than only crops in the agricultural landscape to fulfill their foraging requirements. High pollen diversity is likely to promote colony performance (Hass et al., 2019). Furthermore, bumble bees potentially pollinate not only crops but also many wild plant species. Interestingly, we also detected a large number of sequences derived from fungi (**Figures 6B,D**), which may inhabit flowers (Keller et al., 2015).

We like to mention that the bumble bee samples used for this comparison of sequencing methods are part of a larger study that investigates pollen resource usage of bumble bees in more detail, including a comparison to honey bee foraging and with respect to landscape parameters (Bänsch et al., submitted). The primary focus of the study presented here is the comparison of the applicability of third-generation nanopore sequencing in contrast to established next-(second-)generation sequencing methods. Obviously, both technologies have their strengths and weaknesses. While MinION and NextSeq perform comparably well when querying against an ITS2-specific sequence database, shorter genetic markers still benefit from the higher accuracy of next-generation sequencing.

CONCLUSION

The goal of our study is to compare polymerase (Illumina NextSeq) and nanopore (Oxford Nanopore Technology MinION) generated sequence reads for the assignment of pollen DNA to plant genera. Illumina reads have the advantage of higher quality scores. In contrast, the Nanopore sequencing technology yields longer read lengths. Starting with ITS2-amplicons, we employed two different assignment approaches: (a) BLASTing against a Lower Saxony specific ITS2-sequence database (created within this study) and (b) querying against a k-mer genome sequence database with Kraken2. For (a) the results are comparable: the lower sequence quality is compensated by the read length. For (b) there are two observations striking: (i) the identification of "bycatch" depicted as result of the more extensive database and (ii) the higher amount of assigned taxonomic units on genus level despite the overall smaller read dataset, most likely reflecting the error prone nature of nanopore reads.

In conclusion, we demonstrate the applicability of MinION nanopore sequencing analyzing the dietary diary of bumble bee. Sequence read processing with open software tools and standard parameters yield results close to established next-generation sequencing.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found at bioproject accession PRJNA593728 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA593728). Under Github repository https://github.com/awkologist/BumbleBeeDietaryDiary the ITS2 database used in this study is freely accessible.

AUTHOR CONTRIBUTIONS

SB, CW, and RW conceived the study. SB conducted the field work. SB, BB, RL, and LP performed the DNA extraction and sequencing. RL, LP, and RW analyzed the data. RL and RW wrote the manuscript. All authors reviewed and revised the manuscript.

FUNDING

RL acknowledges funding through European Social Fund (ESF), Ph.D. Scholarship grant number 100316182. SB acknowledges her Ph.D. scholarship from the German Federal Environmental Foundation (Deutsche Bundesstiftung Umwelt). RW received funding from the Saxonian Ministry of Sciences and Arts and the Saxony5 Initiative. CW is grateful for funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 405945293.

ACKNOWLEDGMENTS

The authors like to express their gratitude to the funding agencies. The authors acknowledge support by the Open Access Publication Funds of the Göttingen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00287/full#supplementary-material

REFERENCES

Aizen, M. A., Garibaldi, L. A., Cunningham, S. A., and Klein, A. M. (2008). Long-Term global trends in crop yield and production reveal no current pollination shortage but increasing pollinator dependency. *Curr. Biol.* 18, 1572–1575. doi: 10.1016/j.cub.2008.08.066

Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J., and Förster, F. (2015). ITS2 Database V: twice as much. *Mol. Biol. Evol.* 32, 3030–3032. doi: 10.1093/molbev/msv174

Aronesty, E. (2013). Comparison of sequencing utility programs. *TOBIOIJ* 7, 1–8. doi: 10.2174/1875036201307010001

Belhadj, H., Harzallah, D., Dahamna, S., and Ghadbane, M. (2015). A plausible role for pollen-residing molds in agricultural purposes. *Commun. Agric. Appl. Biol. Sci.* 80, 559–562.

Bell, K. L., de Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., et al. (2016). Pollen DNA barcoding: current applications and future prospects. *Genome* 59, 629–640. doi: 10.1139/gen-2015-0200

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T. L. (2018). Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *Biorxiv [Preprint]* doi: 10.1186/s12859-019-2996-x

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5:e8613. doi: 10.1371/journal.pone.0008613

Coster, W. D., D'Hert, S., Schultz, D. T., Cruts, M., and van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

Danner, N., Molitor, A. M., Schiele, S., Hartel, S., and Steffan-Dewenter, I. (2016). Season and landscape composition affect pollen foraging distances and habitat use of honey bees. *Ecol. Appl.* 26, 1920–1929. doi: 10.1890/15-1840.1

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20, 525–527. doi: 10.1016/j.tplants.2015.06.012

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Garve, E. (2004). *Rote Liste und Florenliste der Farn- und Blütenpflanzen in Niedersachsen und Bremen*. Niedersachsen: Niedersächsisches Landesamt für Ökologie.

Garve, E. (2007). *Verbreitungsatlas der Farn- und Blütenpflanzen in Niedersachsen und Bremen*. Niedersachsen: NLWKN - Naturschutz und Landschaftspflege in Niedersachsen.

Hass, A. L., Brachmann, L., Batáry, P., Clough, Y., Behling, H., and Tscharntke, T. (2019). Maize-dominated landscapes reduce bumblebee colony growth through pollen diversity loss. *J. Appl. Ecol.* 56, 294–304. doi: 10.1111/1365-2664.13296

IPBES (2019). *Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Bonn: IPBES.

Kacániová, M., Pavlicová, S., Hascík, P., Kociubinski, G., Kńazovická, V., Sudzina, M., et al. (2009). Microbial communities in bees, pollen and honey from Slovakia. *Acta Microbiol. Immunol. Hungarica* 56, 285–295. doi: 10.1556/AMicr.56.2009.3.7

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von, der Ohe, K., et al. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biol.* 17, 558–566. doi: 10.1111/plb.12251

Kleijn, D., Winfree, R., Bartomeus, I., Carvalheiro, L. G., Henry, M., Isaacs, R., et al. (2015). Delivery of crop pollination services is an insufficient argument for wild pollinator conservation. *Nat. Commun.* 6:7414. doi: 10.1038/ncomms8414

Krehenwinkel, H., Pomerantz, A., and Prost, S. (2019). Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: current uses and future directions. *Genes* 10:858. doi: 10.3390/genes10110858

Marzinzig, B., Brünjes, L., Biagioni, S., Behling, H., Link, W., and Westphal, C. (2018). Bee pollinators of faba bean (*Vicia faba* L.) differ in their foraging behaviour and pollination efficiency. *Agricult. Ecosyst. Environ.* 264, 24–33. doi: 10.1016/j.agee.2018.05.003

Merget, B., Koetschan, C., Hackl, T., Förster, F., Dandekar, T., Müller, T., et al. (2012). The ITS2 database. *J. Vis Exp.* doi: 10.3791/3806

Nardoni, S., D'Ascenzi, C., Rocchigiani, G., Moretti, V., and Mancianti, F. (2016). Occurrence of moulds from bee pollen in Central Italy–A preliminary study. *Ann. Agricult. Environ. Med.* 23, 103–105. doi: 10.5604/12321966.1196862

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., et al. (2019). Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and reverse metagenomics (RevMet). *Methods Ecol. Evol.* 10, 1690–1701. doi: 10.1111/2041-210X.13265

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., et al. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* 7:giy033. doi: 10.1093/gigascience/giy033

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends Ecol. Evol.* 25, 345–353. doi: 10.1016/j.tree.2010.01.007

Rang, F. J., Kloosterman, W. P., Ridder, J., and de. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19:90. doi: 10.1186/s13059-018-1462-9

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., et al. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.* 15:20. doi: 10.1186/s12898-015-0051-y

Steffan-Dewenter, I., and Westphal, C. (2008). Guest editorial: the interplay of pollinator diversity, pollination services and landscape change. *J. Appl. Ecol.* 45, 737–741. doi: 10.1111/j.1365-2664.2008.01483.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050. doi: 10.1111/j.1365-294X.2012.05470.x

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi: 10.1016/j.tig.2018.05.008

Westphal, C., Steffan-Dewenter, I., and Tscharntke, T. (2006). Bumblebees experience landscapes at different spatial scales: possible implications for coexistence. *Oecologia* 149, 289–300. doi: 10.1007/s00442-006-0448-6

White, R., Pellefigues, C., Ronchese, F., Lamiable, O., and Eccles, D. (2017). Investigation of chimeric reads using the MinION. *F1000Research* 6:631. doi: 10.12688/f1000research.11547.2

White, T. J., Bruns, T., Lee, S., and Taylor, J. (1990). "Amplification and direct sequencing of fungal ribosomal rna genes for phylogenetics," in *PCR Protocols*, eds M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J.

White (Amsterdem: Elsevier), 315–322. doi: 10.1016/b978-0-12-372180-8.50042-1

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257. doi: 10.1186/s13059-019-1891-0

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Xu, Y., Lewandowski, K., Lumley, S., Pullan, S., Vipond, R., Carroll, M., et al. (2018). Detection of viral pathogens with multiplex nanopore minion sequencing: be careful with cross-talk. *Front. Microbiol.* 9:2225. doi: 10.3389/fmicb.2018.02225

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., et al. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* 5:e0013102. doi: 10.1371/journal.pone.0013102

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 Plant Science

ORIGINAL RESEARCH
published: 27 March 2020
doi: 10.3389/fpls.2020.00353

Strategies for Effective Use of Genomic Information in Crop Breeding Programs Serving Africa and South Asia

*Nicholas Santantonio[1†], Sikiru Adeniyi Atanda[1,2,3†], Yoseph Beyene[4], Rajeev K. Varshney[5], Michael Olsen[4], Elizabeth Jones[6], Manish Roorkiwal[5], Manje Gowda[4], Chellapilla Bharadwaj[7], Pooran M. Gaur[8], Xuecai Zhang[3], Kate Dreher[3], Claudio Ayala-Hernández[3], Jose Crossa[3], Paulino Pérez-Rodríguez[9], Abhishek Rathore[5], Star Yanxin Gao[6], Susan McCouch[1] and Kelly R. Robbins[1]**

[1] Section of Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, United States, [2] West Africa Center for Crop Improvement (WACCI), University of Ghana, Accra, Ghana, [3] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [4] International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya, [5] Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India, [6] Genomic Open-Source Breeding Informatics Initiative (GOBii) Project, Institute of Biotechnology, Cornell University, Ithaca, NY, United States, [7] Division of Genetics, Indian Agriculture Research Institute (ICAR), New Delhi, India, [8] Research Program - Asia, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India, [9] Colegio de Postgraduados, Mexico, Mexico

Much of the world's population growth will occur in regions where food insecurity is prevalent, with large increases in food demand projected in regions of Africa and South Asia. While improving food security in these regions will require a multi-faceted approach, improved performance of crop varieties in these regions will play a critical role. Current rates of genetic gain in breeding programs serving Africa and South Asia fall below rates achieved in other regions of the world. Given resource constraints, increased genetic gain in these regions cannot be achieved by simply expanding the size of breeding programs. New approaches to breeding are required. The Genomic Open-source Breeding informatics initiative (GOBii) and Excellence in Breeding Platform (EiB) are working with public sector breeding programs to build capacity, develop breeding strategies, and build breeding informatics capabilities to enable routine use of new technologies that can improve the efficiency of breeding programs and increase genetic gains. Simulations evaluating breeding strategies indicate cost-effective implementations of genomic selection (GS) are feasible using relatively small training sets, and proof-of-concept implementations have been validated in the International Maize and Wheat Improvement Center (CIMMYT) maize breeding program. Progress on GOBii, EiB, and implementation of GS in CIMMYT and International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) breeding programs are discussed, as well as strategies for routine implementation of GS in breeding programs serving Africa and South Asia.

Keywords: genomic selection, genomic prediction, breeding informatics, breeding scheme optimization, plant breeding, trial design

OPEN ACCESS

Edited by:
Mary-Ann Blätke, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
Jorge E. Mayer, Ag RD&IP Consult P/L, Australia
Julio Isidro Sanchez, University College Dublin, Ireland
Uche Godfrey Okeke, Bayer CropScience, United States

***Correspondence:**
Kelly R. Robbins
krr73@cornell.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *22 October 2019*
Accepted: *10 March 2020*
Published: *27 March 2020*

Citation:
Santantonio N, Atanda SA, Beyene Y, Varshney RK, Olsen M, Jones E, Roorkiwal M, Gowda M, Bharadwaj C, Gaur PM, Zhang X, Dreher K, Ayala-Hernández C, Crossa J, Pérez-Rodríguez P, Rathore A, Gao SY, McCouch S and Robbins KR (2020) Strategies for Effective Use of Genomic Information in Crop Breeding Programs Serving Africa and South Asia. Front. Plant Sci. 11:353. doi: 10.3389/fpls.2020.00353

INTRODUCTION

Crop improvement through plant breeding is a process of continuous genetic improvement through selection and recombination of superior lines. The response to selection, or rate of genetic gain, is dependent on multiple factors, expressed in the "breeder's equation,"

$$R = \frac{ir\sigma_a}{L} \quad (1)$$

where R is the response to selection per year, i is the selection intensity, r is the accuracy of selection, σ_a is the additive genetic standard deviation for the trait of interest, and L is the generation interval (Lush, 1936).

Assuming that breeding objectives, selection criteria, available germplasm, and target environments are well defined, the success of a breeding program is largely dependent on the optimal use of available resources to maximize the response to selection (Rutkoski, 2019). Effective breeding programs must re-evaluate breeding strategies as technology, environments, access to germplasm, and consumer needs are constantly changing. While all the aforementioned factors are critical, the ability to identify and effectively implement new technologies can be challenging. This is especially true for publicly funded breeding programs in Africa and South Asia, where resource and infrastructure limitations make the adoption of new technologies particularly challenging. The need to overcome these limitations and improve the effectiveness of breeding programs is urgent, given the historically low rates of genetic gains in many programs serving Africa and South Asia (Godfray et al., 2010; Cobb et al., 2019), expected population growth (Alexandratos and Bruinsma, 2012), and the potential impacts of climate change on crop production (Ritchie et al., 2018).

To achieve rates of genetic gain in crop improvement needed to strengthen and stabilize food security, modern technologies must be adopted and efficiently implemented. One promising approach is genomic selection (GS), where the performance of new lines is predicted based on genome-wide information (Meuwissen et al., 2001). Multiple studies have shown the potential of this methodology to increase the rates of genetic gain in plant breeding programs (Heffner et al., 2009; Beyene et al., 2015; Gaynor et al., 2017; Crossa et al., 2017; Rutkoski et al., 2017), often through the reduction in cycle time, L. However, despite compelling evidence of the potential gains from GS and widespread adoption in animal breeding, public sector plant breeding programs have been slow to routinely adopt GS. Adoption of GS in animal breeding applications benefited from the fact that the use of genomic Best Linear Unbiased Predictors (GBLUP) (VanRaden, 2008) and single-step GBLUP (Legarra et al., 2014) enabled GS implementations that were straightforward extensions of existing breeding approaches. In contrast, optimal implementations of GS in plant breeding programs represent a significant change in how breeding data is analyzed, how breeding decisions are made, and how breeding pipelines are designed. The costs and challenges of large-scale implementation of genomic selection in public sector breeding programs have been a significant barrier to routine implementation despite the potential for significant increases in genetic gain.

A typical inbred or hybrid plant breeding program has this basic structure: (i) selection of parents for crossing, (ii) selfing or use of doubled haploid technology (DH) to achieve the desired level of homozygosity, and (iii) multi-stage field trials of selection candidates (inbred lines or testcross hybrids) to identify best lines or hybrids for release and commercialization as varieties. We generalize this structure as a variety development pipeline (VDP, e.g., Figure 4 of Cooper et al., 2014). A typical VDP evaluates progeny lines in the field for several growing seasons, advancing the best lines at the end of each season, with smaller numbers of lines being tested in more environments in each successive season. Lines that are deemed successful in advanced trials are candidates for variety release and are typically recycled as parents into the breeding program. This approach to breeding takes advantage of the ability to produce inbred or testcross hybrid seed in large quantities which is then extensively evaluated in the field. In this approach, decisions to recycle lines as new parents are made using extensive, but costly, phenotypic data, often with lines treated as independent factor levels in the analysis. This approach produces accurate (r, Equation 1) estimates of line performance but significantly increases generation interval (L, Equation 1) due to the multiple years of testing. While simulation studies demonstrate that a rapid-cycle recurrent GS approach may ultimately provide the largest increases in genetic gains (Gaynor et al., 2017; Gorjanc et al., 2018; Rembe et al., 2019), it is not a practical initial implementation of GS in a plant breeding program. Rapid cycle approaches require relatively large training sets that must be routinely updated to maintain prediction accuracy and breeding decisions must be made using less accurate estimates of the line performance, often without observing the line in replicated trials (Crossa et al., 2010; Hickey et al., 2014; Schopp et al., 2017; Gorjanc et al., 2018). This represents a significant change in how breeding decisions are made and requires significant investments for training set development. Both of these factors can limit adoption of GS, especially in resource limited breeding programs, and these factors need to be considered when developing a strategy for implementation of GS.

Large scale adoption of GS will require optimizing breeding strategies while accounting for costs, ease of implementation, and potential impacts on operation efficiency and genetic gain. Ideally, training data for a rapid-cycle recurrent selection approach would be sourced from the breeding program's VDP. So, regardless of the ultimate end goal and long-term GS strategy, the first step in GS implementation is to routinely genotype lines entering the VDP. For sustainability and routine adoption, this needs to be done without significantly expanding breeding budgets. This requires rethinking how early-stage testing is done in a breeding program. Several approaches have been proposed for incorporating GS in VDPs (Bernardo and Yu, 2007; Cooper et al., 2014; Jacobson et al., 2014; Gaynor et al., 2017; Jarquín et al., 2017; Sukumaran et al., 2018). When evaluating optimal approaches for breeding programs with little or no historical data to train prediction models, strategies that achieve good prediction accuracy from small training sets are critical.

Identifying cost-effective approaches to routinely genotype lines entering the VDP is a critical first step. However, implementation also requires operational capabilities to sample, genotype and generate genomic predictions on a tight turn-around schedule. To do this effectively at scale, advanced breeding informatics systems that include biometrical and quantitative genetics, as well as bioinformatics, are needed. Breeding informatics systems require significant and sustained investment in foundational technologies and computational infrastructure. Fortunately, recent funding initiatives have begun to provide the resources needed to build the foundational capabilities required to modernize and improve the efficiency of public sector breeding programs. The Genomic Open-source Breeding informatics initiative (GOBii)[1] is one such funding initiative with the goal of building the information systems needed for routine application of genomic technologies to improve efficiency of breeding programs targeting crop improvement in Africa and South Asia. In addition to the project's focus on genomic technologies, GOBii is also partnering with other open-source breeding informatics initiatives as part of the Excellence in Breeding (EiB)[2] platform. EiB is being developed as a "complete platform" or set of interconnected tools and strategies designed to increase the efficiency of breeding programs through the adoption of modern technologies and optimal use of breeding resources.

To examine potential approaches for GS implementation, proof-of-concept studies were conducted by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) and International Maize and Wheat Improvement Center (CIMMYT) in chickpea and maize, respectively. To examine optimal strategies for routine implementation of GS of lines entering the VDP, three approaches were compared: (i) the development of a dedicated training set (DTS) in parallel to the VDP to serve as a starting point for GS implementation, (ii) splitting full-sib families for training and prediction (FSTS), and (iii) the use of incomplete block designs across environments, or sparse testing (ST), to obtain good prediction accuracies while reducing plot numbers to offset the cost of genotyping. Here we present the results from these studies and discuss strategies for phased implementation of GS in public sector breeding programs. We also highlight breeding informatics capabilities being developed to enable large-scale implementation of genomic breeding strategies.

MATERIALS AND METHODS

Plant Materials

The two datasets from ICRISAT and CIMMYT are described in detail in Roorkiwal et al. (2016, 2018) and Beyene et al. (2019), respectively. Briefly, the chickpea data consists of 315 lines from two distinct chickpea seed-types, Kabuli (n = 153) and Desi (n = 162), evaluated under rainfed and irrigated regimes in a randomized complete block design with three replicates. All lines were previously genotyped with 2,598 DArT markers (see Roorkiwal et al., 2016 for details). To highlight two contrasting environments, only the rainfed and irrigated environments at ICRISAT from 2013 and 2014 were included in all analyses of the chickpea data.

The maize dataset consists of 849 double haploid (DH) lines from 13 bi-parental families out of the CIMMYT Africa maize breeding program. For demonstration, the three families with the largest family sizes were used for this study (pedigree: CML312/LPS-F64, CML442/LPS-F64, CML536/LPS-F64; size: 91, 108, and 88, respectively). Each DH line was testcrossed to a single tester, and the testcrosses were evaluated in an alpha-lattice incomplete block design with two replications planted in the rainy season with supplemental irrigation as needed in both Kiboko and Kakamega, Kenya, as well as under managed drought conditions during the dry season in Kiboko. The DH lines were genotyped with 9,155 dominant repeat Amplification Sequencing (rAmpSeq) markers at Cornell Life Science Core Laboratory Center, Ithaca, NY, United States (Buckler et al., 2016). Markers were filtered for a minor allele frequency >0.05 and $<10\%$ missing values, resulting in 6,785 markers for use in GS.

As the chickpea data consisted of fixed lines generated from many parents, FSTS predictions were not appropriate in this case, and only DTS and ST predictions were compared. In contrast, as the maize dataset consisted of DH generated from three bi-parental crosses, and as such, FSTS and ST were the most appropriate comparisons. In each comparison, the same number of individuals (i.e., half of the individuals for each population) were assigned to the training and test sets.

Population Structure

Population structure was evaluated using singular value decomposition of the additive genomic relationship matrix, $\boldsymbol{K} = \mathbf{UDU'}$, where $\mathbf{U}$ is a matrix of eigenvectors and $\mathbf{D}$ is a diagonal matrix of eigenvalues. The first two eigenvectors multiplied by their respective eigenvalues were plotted against each other to form a principal component (PC) plot. The proportion of variance explained by each PC is defined as $\mathbf{D}_{ii}/\mathrm{tr}(\mathbf{D})$, where tr() is the trace.

Prediction Model

An unstructured univariate genotype by environment interaction model was used to estimate genetic correlations across environments. This can be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is the vector of a phenotype in each environment, $\mathbf{X}$ is the design matrix for the vector $\boldsymbol{\beta}$ of fixed environmental effects, $\mathbf{Z}$ is an incidence matrix linking observations in $\mathbf{y}$ to individuals, $\mathbf{u}$ is the vector of genetic values and $\mathbf{e}$ is the vector of residuals. The random effects were both considered centered multivariate normal such that $\mathrm{E}[\mathbf{u}] = \mathrm{E}[\mathbf{e}] = 0$ and

$$\mathrm{Var}\left(\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{G}\otimes\mathbf{K} & \mathbf{0}\\ \mathbf{0} & \mathrm{diag}(\sigma_i^2)\end{bmatrix}, \tag{3}$$

[1] www.gobiiproject.org

[2] www.excellenceinbreeding.org

where **G** is the genetic covariance of environments (which must be estimated), and **K** is the additive genomic relationship of individuals, calculated from genetic markers (method I, VanRaden, 2008). Residual variances were considered independent and identically distributed within environment but allowed to differ across environments. Models were fit using the average information algorithm of ASReml (Gilmour et al., 1995; Gilmour, 1997).

Fixed effect values, or Best Linear Unbiased Estimates (BLUEs), were used as true estimated breeding values (EBVs) to compare to the Best Linear Unbiased Predictors (BLUPs), or genomic estimated breeding values (GEBVs). These values were computed using the above model, but allowing **u** to be fixed instead of random, with all observed records included.

Model Validation

For the Desi and Kabuli comparisons, genomic prediction accuracy was assessed using k-fold cross-validation with 10-fold, where records for a random 10th of the individuals were removed (or masked) from the dataset for each fold. Each fold was predicted before the accuracy was calculated as the Pearson correlation between the all predicted BLUPs and the observed BLUEs. The average accuracy across 10 replicates was used as the estimate of genomic prediction accuracy. This was accomplished among both seed-types, within each seed-type, and across seed-types.

For DTS prediction, the population was randomly split into two sets. Phenotypic records from individuals in one set were removed in both environments before fitting the prediction model with records from the remaining set to predict GEBVs for the missing individuals in both environments. Prediction accuracy of unobserved genotype-environment combinations was then determined using the Pearson correlation of the predicted BLUPs to the observed BLUEs either separately by environment, or by combining predictions across environments. This process was repeated 10 times and averaged to produce an estimate of prediction accuracy.

Similar to DTS prediction, FSTS prediction was accomplished by removing phenotypic records from a random half of the lines within each bi-parental family in all three environments. The remaining individuals were used to fit the prediction model and predict the GEBVs of unobserved individuals in all three environments. Results from ten replicates were averaged to estimate prediction accuracy.

Genomic prediction accuracy of ST was determined by again randomly splitting the individuals into two equal sized sets. For ST in chickpea, phenotypic records of one half were removed in the rainfed environment while the records of the other half were removed in the irrigated environment. For ST in the maize dataset, half of the individuals within each family were removed from Kiboko, then further split in half and removed from either Kakamega or Kiboko Drought, along with an additional quarter from the remaining set (see **Figure 1**). Prediction accuracy of unobserved genotype-environment combinations were then determined using the Pearson correlation to the observed BLUEs either across or within the environment. The mean accuracy and standard deviations of replicates for DTS, FSTS and ST can be found in **Supplementary Tables 1**, **2**.

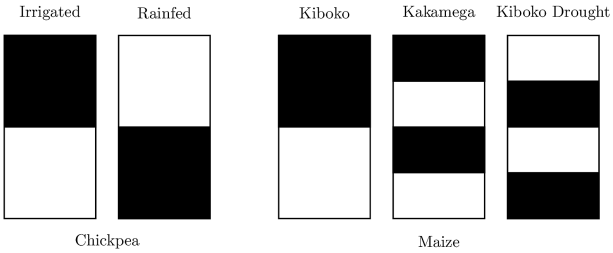

FIGURE 1 | Representation of sparse prediction scheme in maize and chickpea. Black represents records present in the model fit for individuals in each environment, while white represents removed (i.e., missing) records.

RESULTS

Population structures for the chickpea and maize datasets can be found in **Figure 2**. The principal component plot for the maize dataset shows clustering by population but there is a significant overlap between populations. This is not surprising as the maize dataset consists of half-sibs from multiple populations. In contrast, the chickpea data shows two distinct clusters representing Kabuli and Desi lines, which are both genetically and phenotypically distinct.

To determine whether these two chickpea groups should be combined for training and prediction, cross-validation was conducted among and within each group. Prediction across seed-types was also accomplished to determine if the allele frequency and linkage disequilibrium (LD) pattern is sufficiently shared between seed-types. Results from the cross-validation results are found in **Table 1** and **Figure 2**. High cross-validation accuracies were achieved using the combined dataset, containing both Desi and Kabuli lines in both the training and validation sets, in agreement with Roorkiwal et al. (2016, 2018), however, almost complete loss of predictive ability was observed when one seed-type was used to predict the other (**Table 1**). Training and validation sets containing only one seed-type were generally less accurate at predicting performance of that seed-type, as compared with training and validation sets containing both seed-types (**Table 1**).

To determine whether the high prediction accuracies seen using Desi and Kabuli in both training and validation sets were due to the prediction of group differences between Desi and Kabuli, or due to predictions of phenotype variation within seed-type, we then compared (1) single seed-type training sets to predict phenotypes for the same seed-type, with (2) both seed-types to predict phenotypes for a single seed-type. Higher prediction accuracies were generally observed when the training population was consisted of a single seed-type (**Figure 3**).

Genetic correlations between environments vary across traits and range from moderate to high (**Table 2**). Results from cross-validation comparing sparse testing to prediction using historical information can be found in **Figure 4**. Results show that

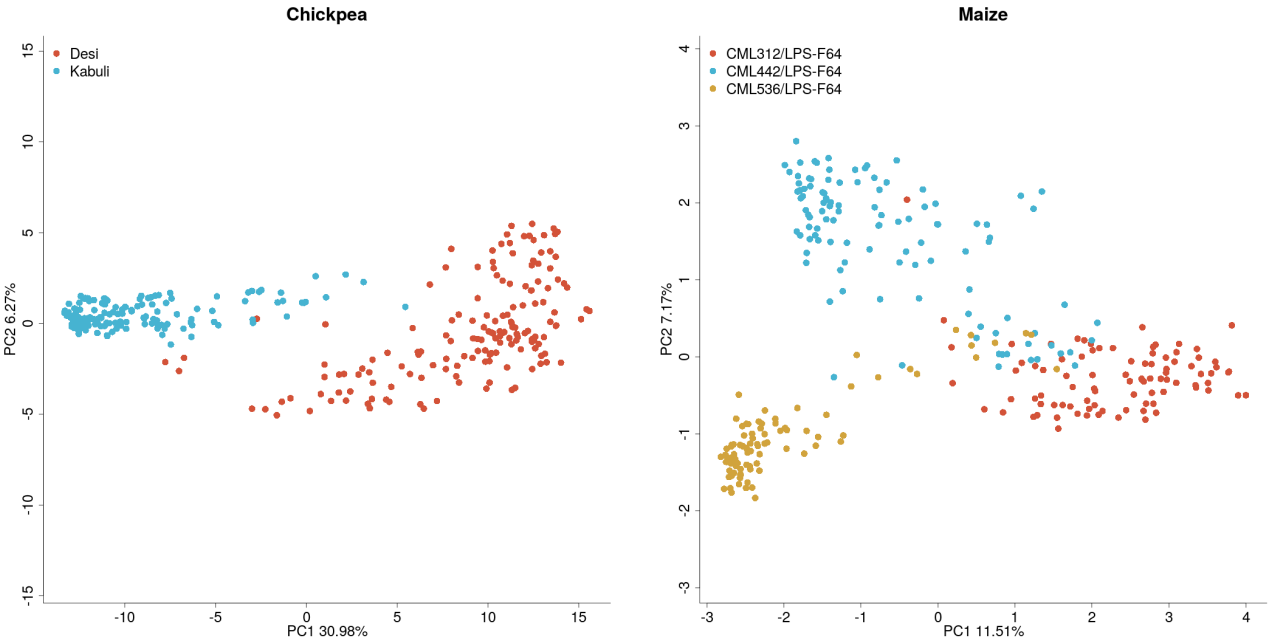

FIGURE 2 | Plots of the first two principal components of the additive genomic relationship matrix for maize and chickpea populations.

TABLE 1 | Genomic prediction accuracies for chickpea GEBVs across environments with various training and test sets estimated using 10-fold cross-validation.

	All predict all	Desi predict Desi	Kabuli predict Kabuli	Desi predict Kabuli	Kabuli predict Desi
Seed yield	0.48 (0.015)[a]	0.26 (0.029)	0.25 (0.020)	0.08[b]	0.04[c]
Seed weight	0.92 (0.002)	0.76 (0.012)	0.74 (0.014)	0.20	0.58
Biomass	0.50 (0.013)	0.39 (0.019)	0.26 (0.026)	0.11	0.16
Plant height	0.65 (0.011)	0.75 (0.010)	0.42 (0.038)	−0.13	0.16
Days to flower	0.68 (0.007)	0.63 (0.016)	0.56 (0.031)	−0.34	0.07
Days to maturity	0.70 (0.003)	0.53 (0.021)	0.53 (0.038)	−0.16	0.09

[a]*Mean prediction accuracy of the Pearson correlation between unobserved BLUPs and observed BLUEs. Standard deviation of ten replicates is shown in parentheses.* [b]*All Desi lines used to predict all Kabuli lines (no cross-validation).* [c]*All Kabuli lines used to predict all Desi lines (no cross-validation).*

across traits, the sparse testing approach consistently achieves prediction accuracies that are as good or higher, which agrees with similar studies in wheat (Jarquín et al., 2017; Sukumaran et al., 2018). Unsurprisingly, the relative improvement in performance increases for traits with higher genetic correlations across environments.

DISCUSSION

Strategy for Phased Implementation

Routine implementation of genomic information represents significant changes in the way plant breeding programs operate and how breeding decisions are made. To facilitate routine implementation, we recommend a phased implementation strategy (**Figure 5**). In Phase 1 the goal should be to establish informatics capabilities to successfully implement GS and optimize trial designs, such as ST and FSTS, to build appropriate training datasets in a cost-effective manner. While the focus of this paper is the routine implementation of GS, it should be noted that routine genotyping of all entries in the VDP will immediately enable genetic quality control and pedigree verification which can improve the overall efficiency of the breeding pipeline by identifying errors early in the screening process. Once the accuracies of genomic prediction models are validated in a breeding program, Phase 2 of implementation should focus on increasing selection intensity in the early stages of the VDP, reducing the number of seasons in which varieties are tested prior to release and recycling lines as new parents earlier in the testing process. Finally, Phase 3 would focus on the implementation of rapid-cycle recurrent selection to reduce generation intervals towards the biological limits of the species. The proposed phases account for the dependencies and logistical complexities of implementation, as well as the size of the training set needed to maintain accurate predictions. Phase 1 assumes a breeding program is starting with very little combined genotypic and phenotypic data to train prediction models. Given that most public sector breeding programs in Africa and South Asia have yet to initiate routine genotyping of lines entering the VDP, the key first step is to implement capabilities and cost-effective

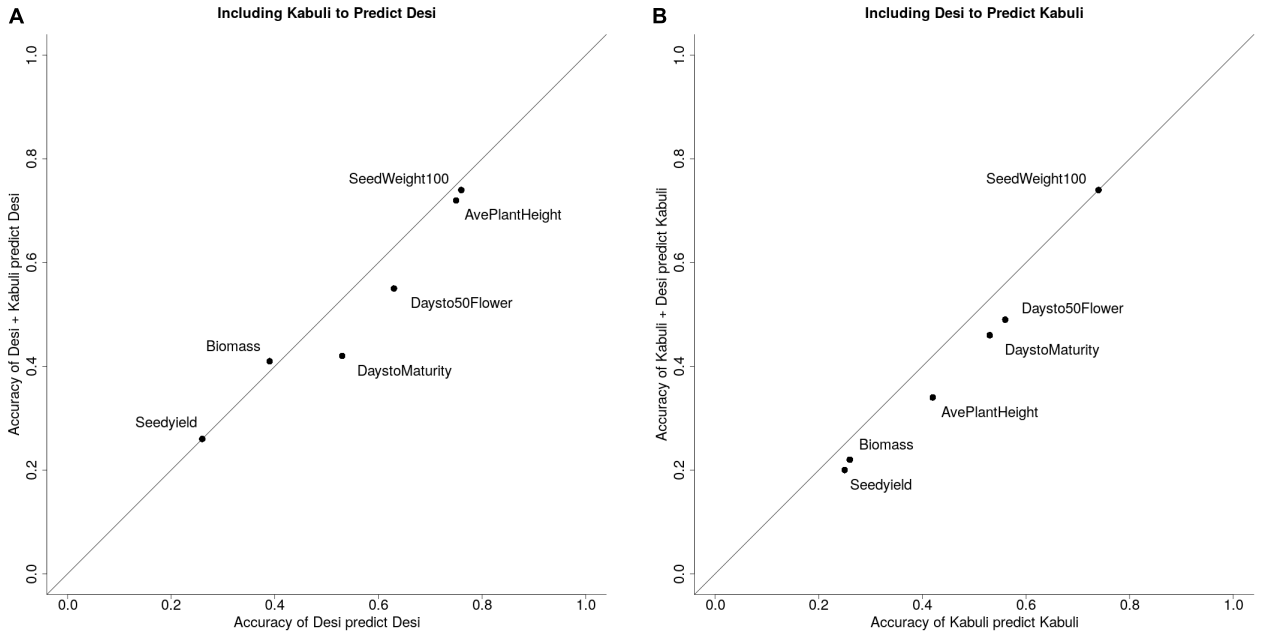

FIGURE 3 | Prediction accuracies for Desi and Kabuli with different structures for training and validation sets. **(A)** Prediction accuracy of using only Kabuli lines to predict Kabuli lines vs. using both Desi and Kabuli lines to predict Kabuli lines. **(B)** Prediction accuracy of using only Desi lines to predict Desi lines vs. using both Kabuli and Desi lines to predict Desi lines.

strategies to routinely genotype lines entering early-stage testing and generate accurate predictions with limited training data. This is a key focus of this study and of projects like GOBii and the High Throughput Genotyping (HTPG) project, both funded by the Bill and Melinda Gates foundation, which are working to increase genetic gains and improve the efficiency of breeding programs serving Africa and South Asia.

The data used in this study were collected on crops with different breeding approaches and different initial strategies for building training sets and applying GS in early-stage trials. These contrasting crops make for an interesting dataset for testing widely applicable strategies for the initial adoption of GS approaches. The chickpea training set was developed to represent the full diversity of ICRISAT chickpea breeding programs, both Kabuli and Desi, for the purpose of predicting the performance of new lines prior to preliminary yield trials. The combined chickpea training set may be good at distinguishing phenotypic differences between the two known groups, but less accurate at discriminating within groups. The inability to predict across seed-types demonstrates that population specific allele frequency and LD patterns appear to be driving the observed prediction accuracy. While using Kabuli lines to predict the performance of Desi, and vice versa, may be viewed as an extreme case, the large decreases in prediction accuracy when compared to the use of Desi to predict Desi and Kabuli to predict Kabuli highlight the importance of building appropriate training sets.

In the chickpea case, the estimates of prediction accuracy using both seed-types were overly optimistic and could have disappointed and discouraged funders of these early GS efforts. Indeed, many reported genomic prediction accuracies are likely upward biased when it comes to selection, as the (unobserved) accuracy of new lines formed from relatively few crosses will not be inflated by the same degree of population structure within the diverse training population. The inability to predict across demonstrates that the two seed-types comprise effectively separate breeding programs and should be treated as such for training population designs in order to provide realistic expectations to funders. It may be prudent to refrain from reporting accuracies in diverse populations, instead focusing on the average of within group/family to guide expectations.

It has been shown that, when assuming the infinitesimal model, the expected prediction accuracy is a function of population structure, trait heritability, training set size, and the accuracy with which genomic relationships calculated using genetic markers estimate the true genomic relationships at the QTL regions controlling the trait of interest (Goddard, 2009; Daetwyler et al., 2010; Goddard et al., 2011). The latter is a function of both marker density and the number of independent chromosomal segments segregating between the training set and the target set of lines for prediction. Strategies that utilize training sets containing lines closely related to the target lines for prediction reduces the number of independently segregating chromosomal segments, which in turn increases prediction accuracy. When LD is high, as it is within close relatives, small training sets and mid to low-density marker platforms can adequately capture the genetic information required for prediction (Schopp et al., 2017; Brauner et al., 2019). A straightforward approach to ensuring training data is closely related to new lines being developed in the breeding programs is to adopt a dual purpose line development and VDP approach to

building training sets, where the VDP serves the additional purpose of providing training data to continuously update predictive models (Schopp et al., 2017; Brauner et al., 2019). While this certainly isn't a groundbreaking revelation, it does provide a clear target for the initial step in implementing GS in a breeding program: cost-effective genotyping of all lines entering the VDP. The general concept of maintaining reasonably close genetic relationships between the germplasm in the line development program and germplasm in the VDP is an important consideration when balancing the effectiveness of the long-term GS implementation strategy with the need to diversify the germplasm base. Maintaining diversity is important for sustained long-term genetic gain as well as response to evolving breeding targets. While GS shows substantial promise for improving breeding program efficiency, it requires a thoughtful germplasm strategy to maximize long-term effectiveness.

When comparing approaches to initiate a dual purpose VDP, the ST approach consistently outperforms both FSTS and DTS in terms of prediction accuracy. Given the differences in crops and population structure of the training sets in this study, the fact that ST delivered higher prediction accuracies in both cases indicates that it could be a robust strategy across crops and breeding programs. It should be noted that the ST method does necessitate the generation of seed from all lines, where the FSTS does not, however, the amount of seed required is less, presenting the potential for time savings during seed multiplication for inbred crops. In lower throughput programs where seed multiplication occurs in the field, this could allow material to enter the VDP an entire year earlier. However, for hybrid crops, the cost implication of seed multiplication for ST is greater since hybrid crops require testcrossing all candidates. The tradeoffs between cost and accuracy need to be carefully considered when considering implementation strategies.

Traits that benefited most from the sparse testing approach were of moderate to high heritability. Traits with low heritability, such as seed yield, also tended to have low genetic correlations across environments. Often, moderate to high heritability traits are under selection in small plot trials during seed multiplication, meaning sparse testing may not be as advantageous for these cases as indicated here. More importantly, the observation of all lines in the field, as is done in ST, allows for a breeder to identify and cull lines with other undesirable, but highly heritable traits, before they enter into extensive field trials. Sparse testing also presents opportunities for cross program collaboration, including across countries or continents. If both programs share a marker platform, implementation of germplasm sharing could be expedited by predicting performance in the other program, and exchange of promising materials for the other environment(s). However, this may be limited to programs that already share related materials which can be reliably predicted.

While prediction accuracy is a major factor in determining the best approaches to implement GS, the cost and complexity of implementation must also be considered. For simplicity and ease of comparison, the same number of plots were used in training predictive models for each approach presented here. This does not mean that each approach would have roughly the same cost or the same efficiency in VDP design. The FSTS approach has the advantage of reducing the number of lines for which seed must be produced for yield trial testing as with this approach phenotypic data is not collected on all genotyped lines. The DTS approach enables prediction of new lines prior to the collection of any information on the line itself or on full-sibs, but requires significant initial investment to develop the training set. Thus, it is difficult to envision an implementation that is cost neutral in terms of the total breeding budget. The ST approach combines genomic prediction and advancement decisions into a single analysis. The fact that implementation can be viewed as a change in experimental design is appealing, but it does increase the complexity of models that need to be run to advance lines through the VDP. In an ST approach, genomic relationship matrices need to be calculated for variety trials and used in mixed models for variety advancements. This adds complexity to the traditional advancement process that could quickly overwhelm even a moderate sized breeding program without breeding informatics tools to support the process.

It is important to note that incomplete block designs typically have some explicit genetic overlap, with some lines shared across

TABLE 2a | Plot level heritabilities and genetic correlations across rainfed and irrigated environments for chickpea.

Chickpea		Desi		Kabuli	
		Rainfed	Irrigated	Rainfed	Irrigated
Seed yield	Rainfed	0.38[a]	0.24[b]	0.29	0.1
	Irrigated		0.32		0.16
Seed weight	Rainfed	0.59	0.88	0.65	0.83
	Irrigated		0.76		0.74
Biomass	Rainfed	0.21	0.41	0.27	0.25
	Irrigated		0.28		0.11
Plant height	Rainfed	0.54	0.87	0.42	0.73
	Irrigated		0.64		0.49
Days to flowering	Rainfed	0.51	0.91	0.55	0.97
	Irrigated		0.6		0.67
Days to maturity	Rainfed	0.36	0.82	0.49	0.89
	Irrigated		0.34		0.38

TABLE 2b | Plot level heritabilities and genetic correlations across three environments for maize.

Maize				
		Kiboko	Kakamega	Kiboko drought
Yield	Kiboko	0.30	0.54	0.72
	Kakamega		0.25	0.40
	Kiboko drought			0.30
Moisture	Kiboko	0.05	0.55	0.98
	Kakamega		0.45	0.31
	Kiboko drought			0.19
Plant Height	Kiboko	0.36	0.86	0.97
	Kakamega		0.27	0.77
	Kiboko drought			0.32

[a] *Plot level heritabilities within each environment are represented on the diagonal.*
[b] *The above diagonal is the estimated genetic correlation of environments.*

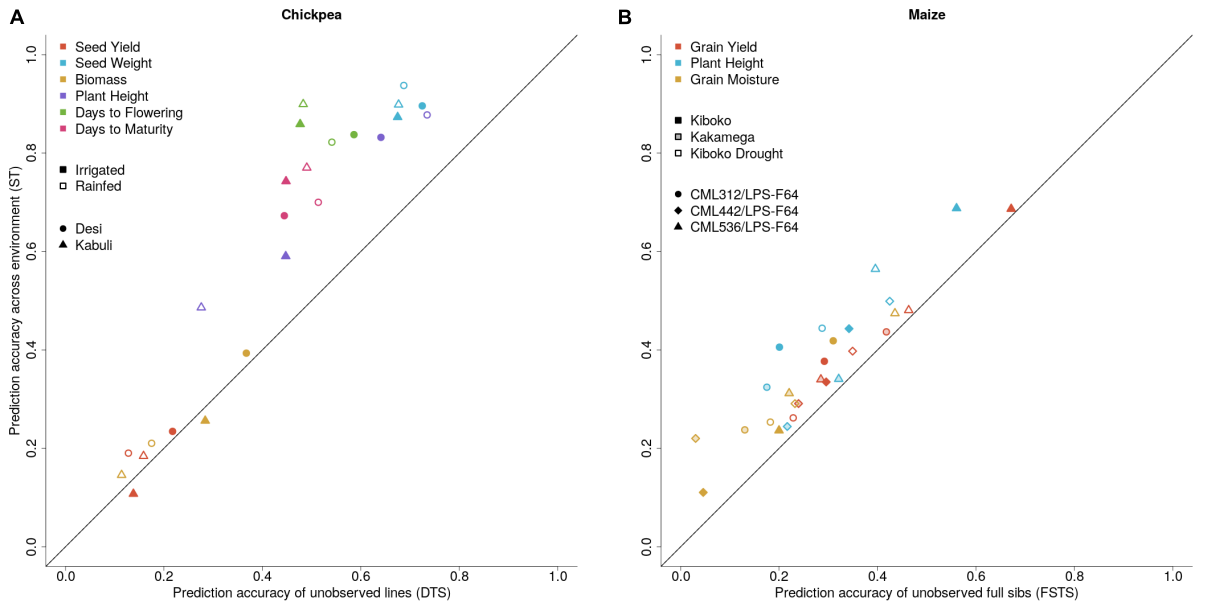

FIGURE 4 | Prediction accuracies for sparse testing (ST) vs. **(A)** dedicated training set (DTS) prediction accuracies in chickpea lines across six traits and two water regimes, and **(B)** Full-sib prediction accuracies (FSTS) in maize across three traits and three environments.

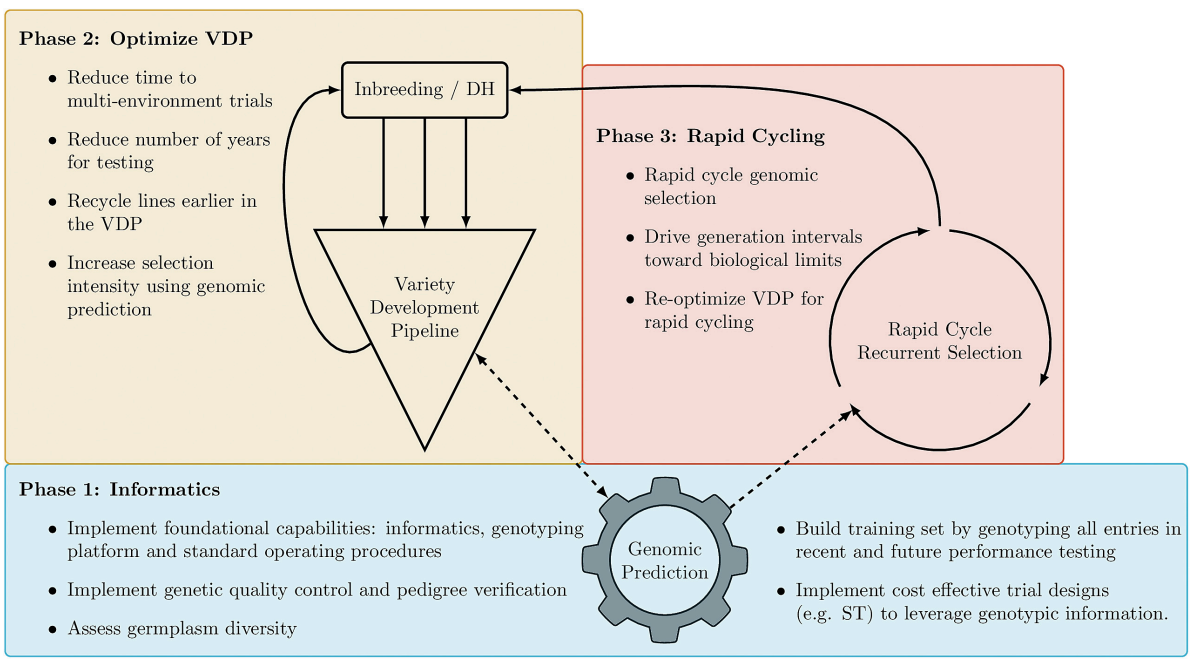

FIGURE 5 | Recommended implementation of genomic information into a breeding program. Phase 1 **(blue)**, Phase 2 **(yellow)**, and Phase 3 **(red)**. Solid lines indicate the flow of genetic materials, while dashed lines indicate the flow of information.

each pair of environments, as is the case here in maize. With this overlap, the genetic correlation of environments can be estimated even when lines are considered independent. When genotypes are available, however, the genetic correlation of environments can be estimated without the need to replicate any lines across environments. These correlations are instead estimated through replication of alleles across environments, as is the case here in chickpea. The ST approach does require estimation of genetic correlations across environments, and this should be taken into account when designing multi-environment trials. Generally, greater levels of genetic overlap will increase the precision of these correlation estimates.

While these results provide some general guidance on promising approaches for initial implementation of routine GS, the optimal implementation strategy will depend on the specifics of each breeding pipeline. The heritability of traits of interest, cost of phenotyping, amount of historical data available for training prediction models, field testing resources, structure of breeding populations, and access to cost-effective genotyping platforms are all factors that will influence decisions about optimal approaches for implementation of GS. Even within a program there may be a need for hybrid approaches given the expected prediction accuracy for a given population using historical data. It is recommended that any breeding program test the potential efficiency of new approaches using simulation prior to implementation. Fortunately, there are freely available simulation packages (Faux et al., 2016; Yabe et al., 2017), and EiB is working directly with public breeding programs in Africa and South Asia to conduct simulations and make recommendations for optimal breeding pipeline designs.

Breeding Informatics

Implementation of any of the approaches examined in this study will require full integration of genomic information into routine breeding decisions, requiring a shift in how data is viewed and handled in a breeding program. The need to build a large training set through a dual purpose VDP means that variety testing trials can no longer be viewed as independent experiments for the identification and advancement of superior varieties. The data collected should be treated as a resource for increasing understanding of breeding germplasm and improving the accuracy of breeding decisions (Spindel and McCouch, 2016). The capability to combine genotypic data with phenotypic data collected across experiments, environments, and seasons will be critical for success. While challenging in and of itself, the narrow timelines between harvesting yield trials and planting nurseries to generate seed for the next season make it infeasible to implement these approaches without effective data management and analytic platforms. To bring genomic information into routine breeding decisions and enable access to valuable data resources, information systems are required to track samples, store genomic and phenotypic information, and implement analysis pipelines to merge data from multiple sources and conduct advanced analytics to guide decision making on tight schedules. In addition, a standardized, low-cost and robust genotyping platform with short turn-around time is essential to delivering high-quality genotyping data in a timely fashion.

To address this critical need, GOBii, EiB, and several other projects are working with public sector breeding programs to build and deploy the foundational capabilities needed to digitize breeding data, support breeding processes and implement GS routinely. Given the size and capacity of many public sector breeding programs, open-source breeding software needs to be both scalable and customizable to meet the needs of diverse crop breeding programs. To accomplish this communities of practice associated with projects like the Breeding Application Program Interface (BrAPI; Selby et al., 2019) and EiB are working to develop best practices and standards to enable interoperability of software being developed across multiple development teams and projects. **Figure 6** represents a high-level, generic architecture focused on the development of web-based breeding software tools. The use of web-based tools enables cloud deployment of

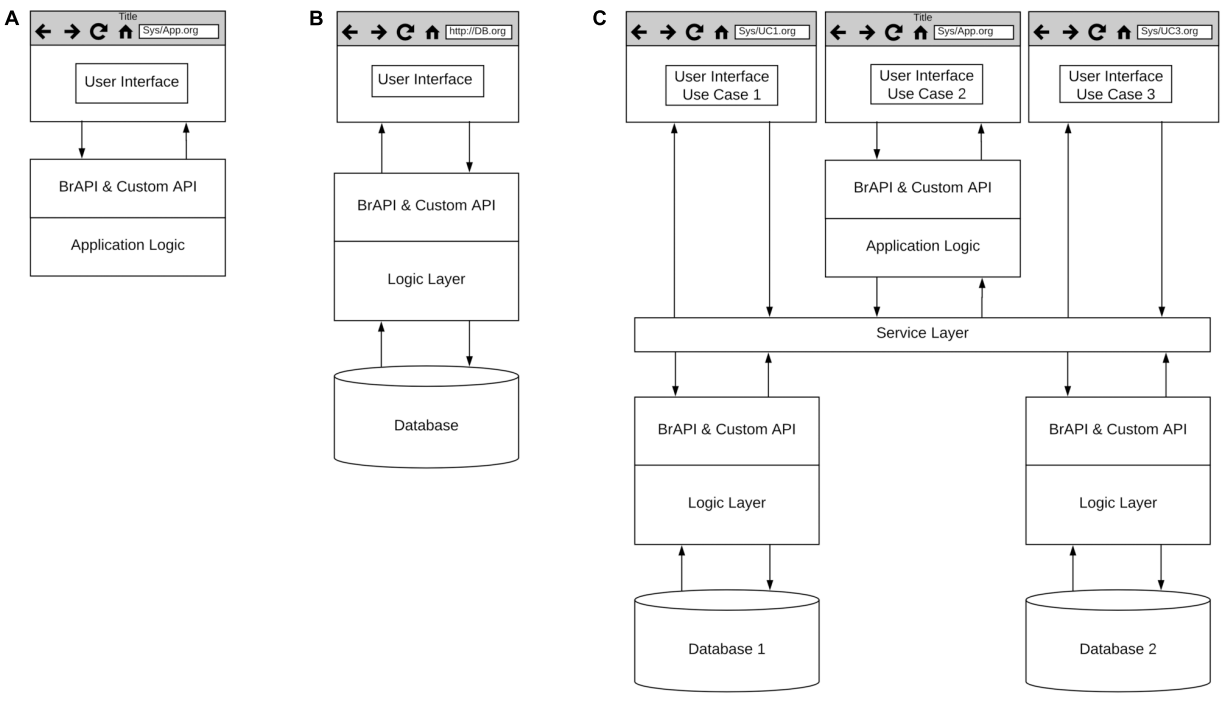

FIGURE 6 | High level architecture for breeding software. **(A)** Applications, **(B)** Databases, and **(C)** Breeding management systems.

complex systems and software as a service (SaaS) for scalability. Common standards for Application Program Interfaces (API), such as BrAPI, will enable customization for a variety of breeding processes. It should be noted that these development efforts are not limited to web-based applications as it is recognized that certain breeding activities will need to be conducted offline.

Several development teams are building applications designed to support specific breeding processes (**Figure 6A**), and several examples of these applications can be found on the BrAPI website[3]. Projects like GOBii are focused on building databases designed to achieve optimal performance with specific data types (**Figure 6B**). The GOBii genomic data management system (GOBii GDM) is designed to store multiple genomic data types and is built on technology that enables fast querying of large genomic datasets (Nti-Addae et al., 2019). The GOBii system utilizes RESTful APIs and the BrAPI standard to enable connections to breeding management systems and breeding analytics pipelines being developed by EiB and other open-source software development projects (Tecle et al., 2014; Ribaut and Ragot, 2019). Finally, several projects such as the EiB Enterprise Breeding System (EBS) and the USDA ARS Breeding Insight[4] are developing systems composed of multiple applications and databases for end-to-end support of breeding processes, leveraging existing breeding software and databases when feasible (**Figure 6C**).

To enable cost-effective genotyping, EiB, in collaboration with the High Throughput Genotyping (HTPG)[5] project, are sourcing genotyping platforms to reduce the cost of mid-density genotyping (1,000–2,000 markers) to a price per line that is comparable to the cost of running a single yield trial plot. Using the HTPG platform, EiB is implementing low-cost genotyping services for public sector breeding programs. Access to these low-cost genotyping services, combined with open-source databases and analytic pipelines greatly reduces barriers to cost-effective implementation of GS strategies and should pave the way for routine use of GS in public sector breeding programs in the near future.

Notably, adoption of new technology demands a skilled workforce. Rapid creation, quality control and turnover of genotypic and phenotypic data will be necessary to make and implement breeding decisions. This will result in many moving parts, and all these steps require a high degree of skill. Many programs will need to adopt a team-oriented approach where expertise is split across many individuals, with enough overlap for effective communication. Future members of plant breeding teams will need skills and expertise outside of what has traditionally been associated with plant breeding. Expertise in database management, machine learning, biometrics, software development, engineering, and operations research will be needed to augment the biology, genetics, and agronomy skills of the team. We acknowledge that building this expertise for every program would be impractical, therefore movement towards regional networks with shared services and expertise will be necessary.

CONCLUSION

There are several barriers to routine implementation of GS at a breeding program scale. These barriers are currently being addressed and we foresee movement towards routine adoption in several public breeding programs. We suggest that breeding programs approach the implementation of GS in a phased approach with the initial step being the routine genotyping of all materials that are evaluated for yield. These materials will be genetically and environmentally close to the materials to be predicted in later stages. We stress that genotyping should be a regular process instead of a series of isolated efforts as is often practiced today. Modification of a traditional variety development pipeline will include implementation of experimental designs that optimize resources allocated to phenotyping and genotyping. Changes in experimental designs and VDP structure should focus on reductions in replications, sparse testing, and faster germplasm turnover. Marker data must be seamlessly integrated with pedigree information, phenotypes, and experimental design to facilitate data processing and analysis for making breeding decisions at a fast turnover rate. Adoption of standardized databases and analysis platforms is necessary to streamline decision making processes. Many of these platforms exist or are currently being constructed, but adoption will be key to successful implementation of GS into the 21st century public breeding program.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to Yoseph Beyene, y.beyene@cgiar.org (maize); Rajeev Varshney, r.k.varshney@cgiar.org (chickpea).

AUTHOR CONTRIBUTIONS

NS analyzed the chickpea data, contributed to the design of the study, and drafting of the manuscript. SA analyzed the maize data, contributed to the design of the study, and the drafting of the manuscript. KR, MO, and EJ contributed to the design of the study and drafting of the manuscript. YB, RV, MG, PG, XZ, CB, and MR contributed to the design of the study and collection of phenotypic and genotypic data. CH, AR, KD, JC, and PP-R contributed to the curation, QC, and analysis of the data. SG and SM contributed to the drafting of the manuscript.

FUNDING

The maize research was supported by the Bill and Melinda Gates Foundation, the Howard G. Buffett Foundations, and the United States Agency for International Development (USAID)

[3]https://brapi.org/brapps

[4]https://www.breedinginsight.org/about

[5]cegsb.icrisat.org/high-throughput-genotyping-project-htpg/

through the Stress Tolerant Maize for Africa (STMA, Grant # OPP1134248), and the CGIAR Research Program MAIZE. The Excellence in Breeding Platform (Grant # OPP1177070) and the Genomic & Open-source Breeding Informatics Initiative (Grant # OPP1093167) are funded by the Bill and Melinda Gates Foundation and CGIAR W1 and W2 funds. The chickpea research was supported by the Department of Science and Technology (DST) Government of India as a part of Australia – India strategic research fund (AISRF) Project and Department of Biotechnology, Govt. of India. This work was carried out as part of the CGIAR Research Program on Grain Legumes and Dryland Cereals.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00353/full#supplementary-material

REFERENCES

Alexandratos, N., and Bruinsma, J. (2012). *World Agriculture Towards 2030/2050: The 2012 Revision. ESA Working paper No. 12-03*. Rome: FAO.

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690

Beyene, Y., Gowda, M., Olsen, M. S., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Front. Plant Sci.* 10:1502. doi: 10.3389/fpls.2019.01502

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., et al. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55, 154–163. doi: 10.2135/cropsci2014.07.0460

Brauner, P. C., Müller, D., Molenaar, W. S., and Melchinger, A. E. (2019). Genomic prediction with multiple biparental families. *Theor. Appl. Genet.* 133, 133–147. doi: 10.1007/s00122-019-03445-7

Buckler, E. S., Ilut, D. C., Wang, X., Kretzschmar, T., Gore, M., and Mitchell, S. E. (2016). rAmpSeq: using repetitive sequences for robust genotyping. *bioRxiv* [Preprint]. doi: 10.1101/096628

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the Breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., et al. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop Pasture Sci.* 65, 311–336. doi: 10.1071/cp14007

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., et al. (2016). AlphaSim: software for breeding program simulation. *Plant Genome* 9, 1–14. doi: 10.3835/plantgenome2016.02.0013

Gaynor, R., Chris, R., Gaynor, C., Gorjanc, G., Bentley, A. R., Ober, E. S., et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386. doi: 10.2135/cropsci2016.09.0742

Gilmour, A. R. (1997). ASREML for testing fixed effects and estimating multiple trait variance components. *Proc. Assoc. Advmt. Anim. Breed Genet.* 12, 386–390.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128, 409–421. doi: 10.1111/j.1439-0388.2011.00964.x

Godfray, H., Charles, J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., et al. (2010). Food security: the challenge of feeding 9 billion people. *Science* 327, 812–818. doi: 10.1126/science.1185383

Gorjanc, G., Gaynor, R. C., and Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Appl. Genet.* 131, 1953–1966. doi: 10.1101/227215

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195

Jacobson, A., Lian, L., Zhong, S., and Bernardo, R. (2014). General combining ability model for genomewide selection in a biparental cross. *Crop Sci.* 54, 895–905. doi: 10.2135/cropsci2013.11.0774

Jarquín, D., da Silva, C. L., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.12.0130

Legarra, A., Christensen, O. F., Aguilar, I., and Misztal, I. (2014). Single step, a general approach for genomic selection. *Livestock Sci.* 166, 54–65. doi: 10.1016/j.livsci.2014.04.029

Lush, J. L. (1936). Genetics and animal breeding*. *J. Hered.* 27, 201–203. doi: 10.1093/oxfordjournals.jhered.a104206

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Nti-Addae, Y., Matthews, D., Ulat, V. J., Syed, R., Sempéré, G., Pétel, A., et al. (2019). Benchmarking database systems for genomic selection implementation. *Database (Oxford)* 2019:baz096. doi: 10.1093/database/baz096

Rembe, M., Zhao, Y., Jiang, Y., and Reif, J. C. (2019). Reciprocal recurrent genomic selection: an attractive tool to leverage hybrid wheat breeding. *Theor. Appl. Genet.* 132, 687–698. doi: 10.1007/s00122-018-3244-x

Ribaut, J.-M., and Ragot, M. (2019). Modernising breeding for Orphan crops: tools, methodologies, and beyond. *Planta* 250, 971–977. doi: 10.1007/s00425-019-03200-8

Ritchie, H., Reay, D. S., and Higgins, P. (2018). Beyond calories: a holistic assessment of the global food system. *Front. Sustain. Food Syst.* 2:57. doi: 10.3389/fsufs.2018.00057

Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., et al. (2018). Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea. *Sci. Rep.* 8:11701. doi: 10.1038/s41598-018-30027-2

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., et al. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* 7:1666. doi: 10.3389/fpls.2016.01666

Rutkoski, J. E. (2019). "A practical guide to genetic gain," in *Advances in Agronomy*, ed. D. L. Sparks (Cambridge, MA: Academic Press Inc.), 217–249. doi: 10.1016/bs.agron.2019.05.001

Rutkoski, J. E., Crain, J., Poland, J., and Sorrells, M. E. (2017). "Genomic selection for small grain improvement," in *Genomic Selection for Crop Improvement*, ed. R. K. Varshney (Cham: Springer), 99–130. doi: 10.1007/978-3-319-63170-7_5

Schopp, P., Müller, D., Wientjes, Y. C. J., and Melchinger, A. E. (2017). Genomic prediction within and across biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. *G3* 7, 3571–3586. doi: 10.1534/g3.117.300076

Selby, P., Abbeloos, R., Backlund, J. E., Salido, M. B., Bauchet, G., Benites-Alfaro, O. E., et al. (2019). BrAPI – an application programming interface for plant breeding applications. *Bioinformatics* 35, 4147–4155. doi: 10.1093/bioinformatics/btz190

Spindel, J. E., and McCouch, S. R. (2016). When more is better: how data sharing would accelerate genomic selection of crop plants. *New Phytol.* 212, 814–826. doi: 10.1111/nph.14174

Sukumaran, S., Jarquin, D., Crossa, J., and Reynolds, M. (2018). Genomic-enabled prediction accuracies increased by modeling genotype × environment interaction in Durum wheat. *Plant Genome* 11, 1–11. doi: 10.3835/plantgenome2017.12.0112

Tecle, I. Y., Edwards, J. D., Menda, N., Egesi, C., Rabbi, I. Y., Kulakow, P., et al. (2014). solGS: a web-based tool for genomic selection. *BMC Bioinform.* 15:398. doi: 10.1186/s12859-014-0398-7

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Yabe, S., Iwata, H., and Jannink, J.-L. (2017). A simple package to script and simulate breeding schemes: the breeding scheme language. *Crop Sci.* 57, 1347–1354. doi: 10.2135/cropsci2016.06.0538

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 Plant Science

ORIGINAL RESEARCH
published: 28 April 2020
doi: 10.3389/fpls.2020.00496

Chromosome-Scale Assembly of Winter Oilseed Rape *Brassica napus*

*HueyTyng Lee[1], Harmeet Singh Chawla[1], Christian Obermeier[1], Felix Dreyer[2], Amine Abbadi[2] and Rod Snowdon[1]**

[1] Department of Plant Breeding, Justus Liebig University Giessen, Giessen, Germany, [2] NPZ Innovation GmbH, Holtsee, Germany

Rapeseed (*Brassica napus*), the second most important oilseed crop globally, originated from an interspecific hybridization between *B. rapa* and *B. oleracea*. After this genome collision, *B. napus* underwent extensive genome restructuring, via homoeologous chromosome exchanges, resulting in widespread segmental deletions and duplications. Illicit pairing among genetically similar homoeologous chromosomes during meiosis is common in recent allopolyploids like *B. napus*, and post-polyploidization restructuring compounds the difficulties of assembling a complex polyploid plant genome. Specifically, genomic rearrangements between highly similar chromosomes are challenging to detect due to the limitation of sequencing read length and ambiguous alignment of reads. Recent advances in long read sequencing technologies provide promising new opportunities to unravel the genome complexities of *B. napus* by encompassing breakpoints of genomic rearrangements with high specificity. Moreover, recent evidence revealed ongoing genomic exchanges in natural *B. napus*, highlighting the need for multiple reference genomes to capture structural variants between accessions. Here we report the first long-read genome assembly of a winter *B. napus* cultivar. We sequenced the German winter oilseed rape accession 'Express 617' using 54.5x of long reads. Short reads, linked reads, optical map data and high-density genetic maps were used to further correct and scaffold the assembly to form pseudochromosomes. The assembled Express 617 genome provides another valuable resource for *Brassica* genomics in understanding the genetic consequences of polyploidization, crop domestication, and breeding of recently-formed crop species.

Keywords: winter oilseed rape, genome assembly, long reads, *Brassica napus*, crop genomics

OPEN ACCESS

Edited by:
Sebastian Beier, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
Mark Timothy Rabanus-Wallace, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Boas Pucker, University of Cambridge, United Kingdom
Jérémy Just, École Normale Supérieure de Lyon, Centre National de la Recherche Scientifique (CNRS), France

***Correspondence:**
Rod Snowdon
rod.snowdon@agrar.uni-giessen.de

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *22 November 2019*
Accepted: *01 April 2020*
Published: *28 April 2020*

Citation:
Lee H, Chawla HS, Obermeier C, Dreyer F, Abbadi A and Snowdon R (2020) Chromosome-Scale Assembly of Winter Oilseed Rape Brassica napus. Front. Plant Sci. 11:496. doi: 10.3389/fpls.2020.00496

INTRODUCTION

Brassica napus subsp. *oleifera*, commonly known as rapeseed or canola, is the second most important oilseed crop globally (Food and Agriculture Organization of the United Nations, 2019). It originated from a natural hybridization event between *B. rapa* (AA, 2n = 2x = 20) and *B. oleracea* (CC, 2n = 2x = 18) no more than 7.5 1000 years ago (Chalhoub et al., 2014). Rapeseed was already widely cultivated in Europe from the 15th to 18th centuries for lamp fuel and soap production (Appelqvist and Ohlson, 1972). Following the introduction of double-low varieties (with low erucic acid and low glucosinolates in the seed) in the 1970s, modern rapeseed/canola varieties today deliver a high-value vegetable oil which can also be used for biodiesel production, while

the extraction meal provides a high quality, protein-rich animal feed (Friedt and Snowdon, 2009). Oilseed rape is also a major component of crop rotations in most cereal-dominated agricultural systems (Friedt et al., 2018).

Brassica napus has an allotetraploid genome composition (AACC, 2n = 4x = 38) (Nagaharu, 1935; Allender and King, 2010). The formation of an allotetraploid involves the challenge to combine subgenomes of distinct species, with individual evolutionary history and epigenetic patterns, into one (Osborn et al., 2003; Comai, 2005). Studies of synthetic allopolyploids and natural neo-allopolyploids showed that the hybridization process of divergent genomes causes instant and prolonged alteration of gene expression, DNA methylation patterns and transposable elements regulation (Salmon et al., 2005; Lukens et al., 2006; Buggs et al., 2010; Chelaifa et al., 2010; Coate et al., 2014; Rigal et al., 2016; Edger et al., 2017). The genomic sequences of allopolyploids are also restructured as a result of illicit pairing of non-homologous chromosomes during meiosis, which encourages homoeologous exchange (HE) events (Gaeta et al., 2007; Xiong et al., 2011). HEs result in the replacement of chromosomal segments of one subgenome with another, and is hypothesized to lead to genome diploidization through fixation of HEs with time (Lysak et al., 2007; Mandáková et al., 2010). In *B. napus*, HEs were revealed through extensive structural rearrangements when the genome of a natural line was compared to the ancestral progenitors (Chalhoub et al., 2014). When compared among seven diverse *B. napus* genotypes, both shared and specific HEs up to a few 100 kb in size were found (Chalhoub et al., 2014), suggesting that HE is an ongoing process in *B. napus*. The mixture of older, fixed HEs and newly-formed HEs explains the wide-spread variations, such as reciprocal (Lombard and Delourme, 2001; Osborn et al., 2003; Piquemal et al., 2005) and non-reciprocal (Udall et al., 2005) translocations, between genotypes. These genotype-specific HEs have been shown to give rise to novel genetic diversities related to important agronomic traits such as flowering time (Pires et al., 2004; Chalhoub et al., 2014; Schiessl et al., 2017), leaf morphology (Osborn et al., 2003; Gaeta et al., 2007), and seed content (Harper et al., 2012; Qian et al., 2016).

The motivation of producing a highly-contiguous *B. napus* genome is clear, particularly from the aspects of breeding research. Genomes of high contiguity enable accurate design of SNP markers to obtain uniquely-mapped probes for marker-assisted selection. A direct comparison of GWAS and genomic selection results between highly-fragmented and chromosomal-scale assemblies of the blueberry genome shows better predictive ability and narrowing of QTL regions (Benevenuto et al., 2019). Similarly, high-quality genomes of the bread wheat (The International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018) and its progenitors (Avni et al., 2017; Luo et al., 2017; Zhao et al., 2017; Ling et al., 2018) have enabled novel gene-to-trait discoveries such as dissection of shattering (Avni et al., 2017) and powdery mildew resistance (Ling et al., 2018). A complete genome assembly also helps fine-tune various decisions in breeding programs, such as target positions for genomic introgressions and identification of potential targets for genome editing CRISPR-Cas9 technologies (Gao, 2018). The genomic characteristics of *B. napus* increase the complexity of studying the genome sequences in three major respects. Firstly, homoeologous regions between subgenomes hamper the genome assembly process due to low sequence specificity. Ambiguities of highly similar sequences are difficult to resolve for assembly algorithms, particularly during the read clustering process (Nagarajan and Pop, 2013). In *B. napus*, reads originating from homoeologous regions cannot always be accurately assigned to individual subgenomes, and subgenomic distinction is further blurred by ongoing HE events. Recent assemblers adopt the *k*-mer binning method to resolve haplotypes using parental genome assemblies (Koren et al., 2018). Since high-quality assemblies of the *B. napus* progenitors are available (Belser et al., 2018), this approach could be plausible however it is nevertheless unable to fully resolve HE events, which interfere with subgenomic separation of *k*-mers. Secondly, as in many other complex crop genome assemblies, the high content of repetitive sequences in *B. napus* interfere with the construction of continuous chromosomes. The two diploid progenitors *B. rapa* and *B. oleracea* are both products of multiple paleopolyploidization events, where large-scale rearrangements occur following divergence from a common ancestor (Parkin et al., 2005). As a result, *B. napus* has potentially accumulated up to 72x multiplication since the origins of angiosperms and about 34.8% of the genome are estimated to be repeats (Chalhoub et al., 2014). Thirdly, the genome assembly of any single cultivar always fails to capture the entire genomic repertoire in a species, hence the need to use a pangenome as reference is recognized in crops (Tao et al., 2019). In oilseed rape, this need is highlighted by the HE-driven variations found between cultivars. Recent evidence shows ongoing HEs even in homozygous cultivars during self-pollination (He et al., 2017), suggesting that the variations between individuals of the same cultivar could be largely underestimated.

Genome assemblies for three cultivars of *B. napus* have been published to date (Chalhoub et al., 2014; Bayer et al., 2017; Sun et al., 2017), with Darmor-*bzh* and Tapidor being the two winter-type genotype represented. The Darmor-*bzh* genome is widely used as a standard reference genome from studies ranging from gene loss (e.g., Hurgobin et al., 2018) to SNP marker-assisted analyses like genome wide association studies (GWAS) (e.g., Gabur et al., 2018). However, all three genome assemblies were constructed prior to the advance of long-read technologies, therefore these assemblies are highly fragmented. To illustrate, the N50 read length of an Oxford Nanxopore MinION single flowcell run today is about 32 kbp (**Supplementary Table S2**), approximately the same size as the N50 contig length in the Darmor-*bzh* assembly (Chalhoub et al., 2014). The long-read technologies, led today by Pacific Biosciences (Eid et al., 2009) and Oxford Nanopore Technologies (Loman et al., 2015), revolutionized genomic research by producing continuous sequences of 10s to 100s of kilobases in length. They are now used to resolve complex and repetitive regions in plant genomes (for example Schmidt et al., 2017; Belser et al., 2018). Long-reads are therefore well-suited to resolve the aforementioned complications in assembling the *B. napus* genome by encompassing HE breakpoints and transposable elements.

Here we report the sequencing and genome assembly of the German winter oilseed rape accession 'Express 617' using 54.5x coverage with Pacific Biosciences long reads. Express 617 is a natural winter oilseed rape accession widely used in many existing mapping populations for linkage analyses of traits such as seed quality (Badani et al., 2006; Stein et al., 2013, 2017), seed yield and yield architecture (Radoev et al., 2008), heterosis (Basunanda et al., 2010) and disease resistance (Obermeier et al., 2013). Short reads, optical map data and genetic maps were used to further correct and scaffold the assembly to form pseudochromosomes. The Express 617 genome is assembled to 925 Mb in size, approximate to the flow cytometry estimation of 1132 Mb (Johnston et al., 2005). The base accuracy and pseudomolecule contiguity were validated using short read libraries, SNP markers and long read alignments. The genome was annotated to contain 12.5% of coding sequences (89857 predicted genes) and 37.5% of repetitive elements. The assembly was also compared to two other published *B. napus* genomes to identify collinear regions. A total of 56 same-chromosome collinear blocks of 488 Mb in size were identified in Express 617 (53%) when compared to the Darmor-*bzh* genome. In comparison, only 230 Mb (25%) of Express 617 are collinear with the ZS11 genome. This long-read genome of *B. napus* is expected to contribute to further understanding of HE in *B. napus* and its role in generation of genetic diversity for quantitative trait expression (Gabur et al., 2019). This assembly expands the genomic repository of *B. napus*, particularly for winter-type accessions, and consequently promotes exploitation of genomics advancement in oilseed rape and canola breeding programs.

TABLE 2 | Gene annotation and evaluation of the Express 617 genome.

	Express 617
Number of genes	89857
Number of transcripts	99481
coreGF	Weighted score: 0.95
	Number of missing coreGFs: 159
Number of BUSCOs found	Total complete: 4358 (94.8%)
	Complete single copy: 866
	Complete duplicated: 3492
	Fragmented: 11
	Missing: 227
Number of transcripts aligned to pan-transcriptome	87012
Number of proteins containing InterPro domains	87951

RESULTS

The Express 617 Genome Assembly, Gene Set, and Repetitive Elements

The total size of Express 617 genome assembly is 925 Mb, where placed pseudochromosomes are 765 and 160 Mb remained as unplaced random scaffolds (**Supplementary Table S3**). As shown in **Table 1**, this genome size is larger than three previously published assemblies, whereas the percentage of N-bases (quantity of gaps) is lowest among all five assemblies. The high contiguity of Express 617 is also reflected in the length of N50 scaffolds (4.8 Mb) prior to pseudochromosome construction.

The genome consists of 12.5% coding sequences, 89857 genes with 99481 transcripts (**Table 2**) and 37.5% repetitive elements (**Supplementary Table S4**). The transcripts have an average length of 1924.8 bp, with an average of 5.22 exons each. Average lengths of intron and exon are 183.3 and 226.5 bp, respectively. A total of 87951 transcripts contain at least one known protein domain that can be found in curated protein databases. As observed in all other plant genomes, the majority of the Express 617 repeats are long terminal repeats (LTRs) (28.3% of all repetitive bases masked), where 22.2% are Gypsy and 16.8% Copia retrotransposons. The non-LTR subclass I and subclass II comprise 4.9 and 13.5%, respectively, while the remaining transposable elements remain uncharacterized (25.5%). Satellites, simple repeats and low complexity sequences make up another 5.2% of all repeat sequences.

Consistent with previous studies (Chalhoub et al., 2014), the chromosomes of subgenome A have higher gene density with lesser repetitive elements when compared to subgenome C (**Supplementary Figure S1**). This is explained by subgenomic dominance, a phenomenon documented in many polyploids such

TABLE 1 | Assembly statistics of the Express 617 genome in comparison with three previously published *B. napus* genome assemblies.

	Darmor-*bzh* v4.1 (Chalhoub et al., 2014)	**Darmor-*bzh* v8.1 (Bayer et al., 2017)**	**ZS11 (Sun et al., 2017)**	**Tapidor (Bayer et al., 2017)**	**Express 617**
Total genome size (Mb)/percentage of Ns	850/13.17%	850/13.16%	976/7.05%	636/5.16%	925/0.09%
Length of pseudochromosomes (Mb)	645	798	854	627	765
Length of unplaced scaffolds (Mb)	204	51	120	8.4	160
Number of scaffolds prior to pseudochromosome construction	20702	-	3460	21280	1632
Length of N50 scaffold prior to pseudochromosome construction (bases)	763688	-	602220	197031	4882293

as cotton (Renny-Byfield et al., 2015) and maize (Schnable et al., 2011), where homoeologous copies of "dispensable" genes are preferentially silenced (Edger et al., 2018).

Evaluation of the Assembly Quality

We took multiple steps to avoid common errors in the assembly process, and then extensively evaluated the results. The correctness of the assembly was evaluated in three ways, (1) base-level errors, (2) large-scale translocations, and (3) completeness of the gene set.

Base-level errors are single nucleotide mutations and short indels that usually arise from the sequencing process. The error rate of raw PacBio long reads was estimated to be up to 15% (Korlach, 2013). Using the alignment of two libraries of Illumina short reads, we assessed the error rate of a subset of PacBio reads (10% of total nucleotides in all reads). By allowing single-end mapping, 7% of the total nucleotides of mapped reads were mismatches, which is half of the maximum estimated error assuming that Illumina reads have near-to-zero sequencing errors (Glenn, 2011). To reduce the effect of long read sequencing errors, we used consensus long reads that were generated by self-alignment. We also incorporated high coverage of short reads during the assembly, as well as post-assembly error correction. To measure the base-level accuracy of the genome, five libraries of Illumina sequencing reads, of which four were used to construct the assembly and one was sequenced independently, were used. A total of 89% of paired-end reads aligned concordantly in the correct direction and insert size, with zero mismatches and gaps (**Supplementary Table S5**).

Large structural error is a primary concern when assembling polyploidy genomes, particularly allopolyploids with frequent HEs like *B. napus* (Samans et al., 2017). These errors could manifest through a few assembly processes, for example (1) wrongly-placed scaffolds during the construction of pseudochromosomes due to non-specific matching to genetic maps, (2) short mate-paired libraries and linked-reads could be unspecific to differentiate between regions of homeoelogous chromosomes, and form wrongly-joined scaffolds, and (3) regions with high density of repetitive elements may form small scaffolds and could be wrongly-placed as described in (1). To evaluate large-scale errors, a combination of SNP markers and long read alignments was used. First, the distribution pattern of gene allelic SNPs in the AC *Brassica* genomics platform (He et al., 2015) generated by the genome-ordered graphical genotypes (GOGGs) method (He and Bancroft, 2018) was manually inspected. The correctness of the assembly was measured by low amount of alternating parental alleles in individual recombinant lines, with the assumption that allelic SNPs segregate across a mapping population while interhomeolog and interparalog SNPs do not. We detect a total of 24 regions of discording allelic patterns indicating putatively incorrect gene order, which could originate from incorrectly placed scaffolds or misjoin of scaffolds (example in **Supplementary Figure S2**). They were labeled as potentially misassembled. To confirm that they were indeed true misjoins and not inaccuracies introduced during GOGG such as ambiguously-mapped orthologs, we used alignments of long reads to the assembly. Long read alignments were generated using PacBio reads which were used to construct the assembly and additionally 17x of Nanopore ONT reads. A total of 86% (562142) of the Nanopore reads and 99% (4328786) of PacBio reads aligned to the assembly. True misjoins were identified by refining the resolution of breakpoints, which are characterized as a huge decrease of mapped reads and an enrichment of split-reads, as seen in **Supplementary Figure S3**. When supported by high coverage of mapped reads for both PacBio and Nanopore technologies, a putatively misassembled region was dismissed, as shown in **Supplementary Figure S4**. Using read coverage as supporting evidence, a total of seven regions, ranging from 123 kb to 3 Mb were identified as misassemblies. All cases of true errors have one or both breakpoints in stretches of Ns. Ns were introduced as gap-fillers during the construction of pseudochromosomes and scaffolding of using linked-reads. Regions with high frequency of Ns therefore symbolizes difficult regions where their local sequence proximity, termed "edges" in an assembly graph (Wick et al., 2015), cannot be resolved. These regions were extracted and retained as unplaced scaffolds. **Supplementary Figure S5** shows the final arrangements of scaffolds based on the genetic versus physical distance of a total of 24469 markers (17478 in Express 617 × R53; 8469 in Express 617 × V8; 12140 in Express 617 × SGDH14) in each pseudochromosome. The pseudochromosomes were also compared to the progenitors *B. rapa* and *B. oleracea* genomes to show that sequence similarities of subgenomes are as expected (**Supplementary Figure S6**).

The completeness of predicted genes were evaluated with a set of well-conserved genes across plant species using PLAZA coreGF (van Bel et al., 2012) and BUSCO (Simão et al., 2015). Out of 2928 core green plants gene families in PLAZA, 2803 were identified in the predicted gene set, therefore obtaining a weighted score of 0.95. BUSCO (v4.0.4) detected 4358 (94.8%) out of 4596 complete orthologous groups within the Brassicales lineage dataset, with 3492 being duplicated and 866 being single copy. In comparison (**Supplementary Figure S7**), version 4.1 of Darmor-*bzh* has 4378 (95.2%) complete BUSCOs, version 8.1 of Darmor-*bzh* has 4379 (95.2%), ZS11 has 4263 (92.7%) and Tapidor has 4162 (90.6%). Additionally, a publicly-available single-ended RNAseq library was used to ballpark the accuracy of annotated introns. A total of 342674 introns were predicted using RNAseq, where 229278 (67%) matched to the introns annotated, 104883 overlapped with predicted gene region, and 8513 were in intergenic regions. This indicates that 62% of introns annotated are supported by external data, which was not used in the annotation pipeline.

Comparison Between Express 617 and Other *B. napus* Genomes

Whole genome alignment between Express 617 and Darmor-*bzh* shows high sequence similarity in all chromosomes (**Figure 1**). The secondary alignments of lower similarity between homeologous chromosomes can also be observed.

To examine the shuffling of chromosomal segments, the gene-level collinearity between genomes, which is defined as the conservation of gene order within syntenic regions

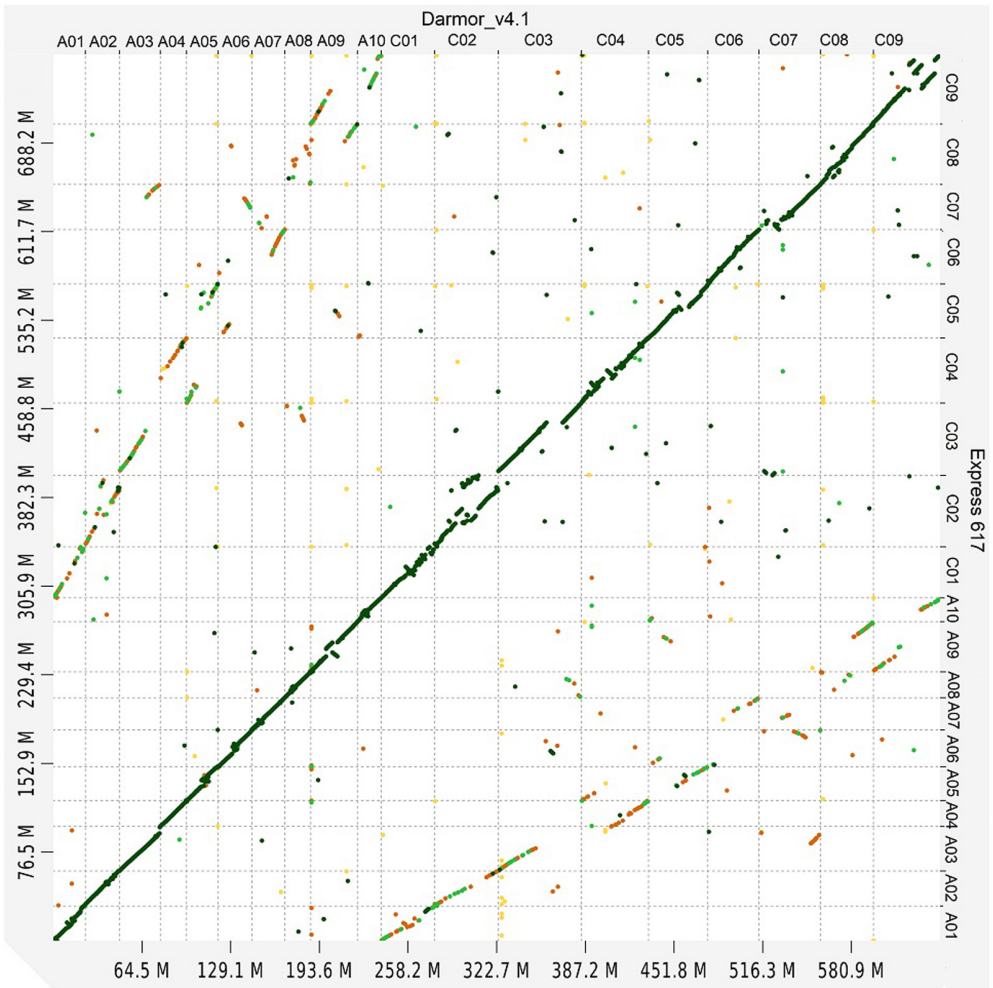

FIGURE 1 | Dot plot comparison between Express 617 and Darmor-*bzh* v4.1 genome. Sequence similarity is color coded from 0 to 1.

(Coghlan et al., 2005), was identified. A total of 120 collinear blocks, linked by 77840 gene pairs, were found between Express 617 and Darmor-*bzh* (**Figure 2A**). Out of 120, 56 blocks linked by 45410 gene pairs correspond to the same chromosomes. These 56 blocks made up 488 Mb (53%) of the Express 617 genome and 425 Mb (57%) of the Darmor-*bzh* genome. In comparison, Express 617 shared 100 collinear blocks (38145 gene pairs) with ZS11 (**Figure 2B**), a spring oilseed rape line, where 56 of the blocks (21982 gene pairs; 230 Mb of Express 617 and 274 Mb in ZS11) are in the same chromosomes.

DISCUSSION

Improved contiguity of Express 617 genome in comparison to other *B. napus* assemblies is evident in the low number of scaffolds, high N50 scaffold length and the low percentage of Ns in total genome size (**Table 1**). This improvement is expected as the all other four are short read assemblies. The Tapidor assembly has the lowest contiguity as it was assembled with about 30x of Illumina short reads and the contigs were placed using SNPs (Bayer et al., 2017). ZS11 and Darmor-*bzh* were constructed with more comprehensive data, including higher coverage of short read sequencing, long range mate-paired reads and BAC-by-BAC approach (Sanger-sequenced for Darmor-*bzh* and Illumina-sequenced for ZS11) (Chalhoub et al., 2014; Sun et al., 2017). For both assemblies the general approach was that first BAC sequences were used to form contigs, gaps were filled with short reads, and then genetic maps were used to place contigs. Since the maximum read length was 100 bp, the assemblies produced are enriched with gaps, as reflected on the percentage of Ns. Better contiguity also means that the intergenic and repeat-rich regions such as centromere are better assembled. This will enhance the development in molecular breeding such as application of transposable element markers (Bhat et al., 2020), following increasing understanding of the role of transposable elements in crops, such as in disease resistance [example in pepper (Kim et al., 2017)], domestication [example in rice (Li et al., 2017)], and adaptations [example in maize (Lai et al., 2017)].

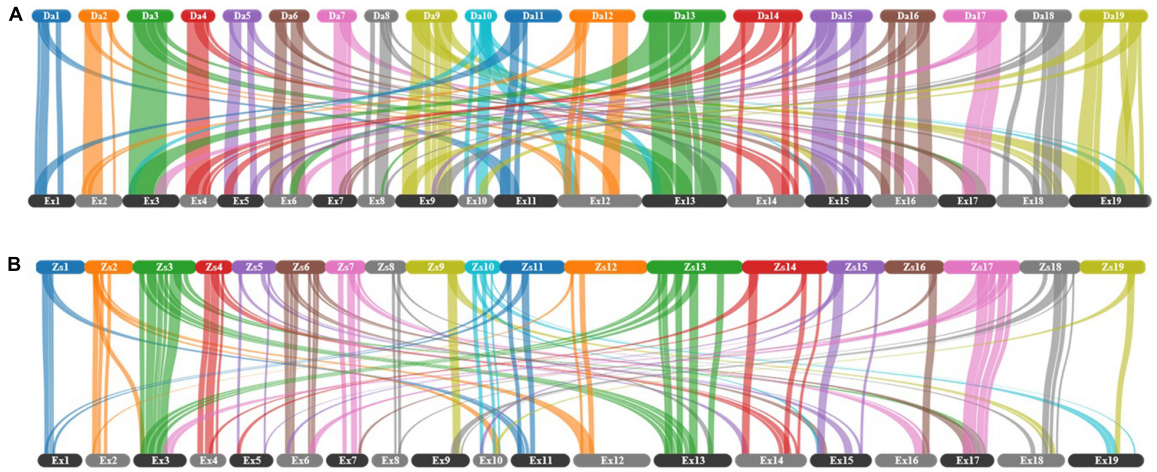

FIGURE 2 | Collinearity between Express 617 and two other *B. napus* assemblies for all chromosomes. **(A)** Darmor-*bzh* v4.1 versus Express 617, **(B)** ZS11 versus Express 617. Collinear blocks are indicated as connecting bars between genomes. The chromosomes were labeled with two letter indicating the cultivar followed by chromosome number, where 1 to 10 corresponds to chromosomes A1 to A10 and 11 to 19 corresponds to chromosomes C1 to C9.

A considerable amount of scaffolds (160 Mb) cannot be placed in the Express 617 pseudochromosomes. Due to the subgenomic similarities and frequency of HEs, a relatively conservative approach was taken to construct this assembly. To avoid false positives and wrong conclusions led by misassemblies, the assembled scaffolds were broken in two independent steps, therefore trading off contiguity for accuracy. Misjoins were broken first during the incorporation of optical map data, and second during GOGG evaluation. Optical maps provide independent long-range evidence for the connection of scaffolds. However, in the circumstances of conflicts between optical map and the assembled sequences, a decision has to be made to resolve conflicts. Since optical maps are not error-free (Jiao et al., 2017), the software *Chimericognizer* (Pan and Lonardi, 2019) used alignments adjacent to the conflicts to estimate the confidence of chimeric sites. A total of 92 scaffolds (out of 1547) in the assembly were identified as chimeric and broken to form 206 scaffolds. Stitching of all the scaffolds was then attempted next using *Novo&Stitch* (Pan et al., 2018). Similarly, using the GOGG approach followed by long read mapping, a total of seven regions were identified as misjoins and breakpoints were cut. Even though the correct chromosome and position of these misassembled blocks can be identified using GOGG patterns, they are of low resolution. In other words, there is no way to accurately determine a breakpoint for insertion of these blocks. These blocks were therefore retained as random scaffolds, with the putative chromosome appended to the scaffold name (**Supplementary Table S3**). Out of all unplaced scaffolds, 38 Mb were assigned to chromosomes and contain 2946 genes, whereas 122 Mb contain 2803 genes, with 34.3% of them being repetitive elements.

The completeness of gene space is one of the ways to evaluate an assembly (Veeckman et al., 2016). Coding sequences made up of 12.5% in Express 617, comparable to 11.9% in Darmor-*bzh* v4.1 (Chalhoub et al., 2014). Both coreGF and BUSCO indicate a 95% completeness of conserved orthologous groups in Express 617 genome. Based only on BUSCO results, this is comparable to Darmor-*bzh*, where 20 more orthologous groups were identified, and more superior to Tapidor and ZS11 (**Supplementary Figure S7**). We postulated that the error rate of long read sequencing affect the accuracy of gene prediction, as observed in human genome assemblies (Watson and Warr, 2019). This could possibly also be reflected on the lower number of confident genes in Express 617 when compared to Darmor-*bzh*. Using Illumina short reads as a benchmark, PacBio raw reads have an overall mapping rate of 71% and an error rate of 7%, whereas the assembly has an overall mapping rate of 99% (perfect mapping rate of 89%) and an error rate of 0.5%. This improvement is largely contributed by the pre-assembly consensus read construction and multiple rounds of short read polishing. We anticipate better polishing softwares, such as sequencing signal-based tools (such as Nanopolish[1]) to resolve the 0.5% uncorrected errors. Nevertheless, we argue that from the perspective of the amount of resources and time used, long read technology has definitely increased the efficiency to produce high quality genomes. For example, the BAC-by-BAC pooled strategy used in Darmor-*bzh* is known to be highly accurate yet expensive and include labor-intensive processes such as fingerprinting clones.

Another possibility for undetected errors to persist in this assembly is that when correlating genetic maps to physical positions, two assumptions were used (1) the genetic maps accurately represent the Express 617 genome, and (2) each marker probe mapped correctly to the chromosomal position of origin. However, even though Express 617 is the common parent of three populations used to generate genetic maps used, there are two populations with parents of synthetic backgrounds (R53 and

[1]https://github.com/jts/nanopolish

V8). Synthetic accessions are known to contain more HEs than non-synthetics (Sharpe et al., 1995; Liu et al., 2014; Rousseau-Gueutin et al., 2017; Hurgobin et al., 2018). For example, a large part of chromosome C02 in R53 was known to be replaced by A02 (Stein et al., 2017). To reduce the manifestations of these HEs in Express 617 pseudochromosomes, weighted priority was given to population of natural lines (Express 617 × SGDH14). The limitation of the second assumption is the specificity of probe mapping. As the length of marker probe is only 50 bp, it could map to multiple positions (8227 out of 44113 of mapped probes are non-unique). Even with uniquely mapped probes, scaffolds could still be wrongly assigned to homeologous or similar regions of non-homeologous chromosomes. Homeologous mappings can be observed in **Supplementary Figure S5**, particularly between A01 and C01, A03 and C03, A09 and C08, and A09 and C08, which are known hotspots for HE events (Chalhoub et al., 2014; Lloyd et al., 2018). Also, the density of markers are not consistent along the chromosomes. To illustrate, the first misjoin in chromosome A01 (position 1846993) detected by GOGG only have adjacent markers at a 58 kbp distance upstream and 139 kb downstream. Since only uniquely-mapped markers were used, repetitive or highly homoeologous regions contribute to large gaps between markers. This potentially explains how the misjoin was formed during the assembly, and how it was not detected with 10x linked reads and optical mapping.

Brassica napus morphotypes cluster into winter, semi-winter and spring growth habits based on SNPs (Diers and Osborn, 1994; Becker et al., 1995; Bus et al., 2011; Gazave et al., 2016; Delourme et al., 2018) and show sequence and copy number variation in flowering time regulatory genes (Schiessl et al., 2017). Collinearity comparisons between the two winter-type cultivars Express 617 and Darmor-*bzh,* and between Express 617 and the Chinese semi-winter cultivar ZS11 reflected this expectation of genetic diversity. However, we nevertheless cannot disregard the influence of assembly quality and completeness in collinearity studies. Regions that are not collinear could arise from true genetic diversity, assembly artifacts such as misassemblies and gap regions, or unidentified genes. Repetitive elements in the genome are likely to be the major contributor of these regions. For example, repeat-masking approach was found to be the main cause of varying number of repeat-containing disease resistance genes in four *B. napus* genomes, instead of true biological variations (Bayer et al., 2018).

MATERIALS AND METHODS

Whole Genome Sequencing of Plant Material

Illumina and Pacific Biotechnologies Sequencing

Approximately 40 g of fresh leaf tissue was collected from an advanced inbred line ($>F_{11}$) of the winter type oilseed rape accession "Express 617." DNA libraries of 350 bp, 450 bp, 2 kbp, 5 kbp, and 10 kbp were constructed and subjected to paired-end sequencing on the Illumina HiSeq 2000 platform. The 20-kb SMRTbell library was prepared using SMRTbell Template Prep Kit 1.0-SPv3, where the qualified high-molecular weight DNA were fragmented to approximately 20 kb, followed by damage repair, end repair and adapter ligation. Size selection was then performed using BluePippin™ Size-Selection System (Sage Science, Beverly, MA, United States). The quality of purified library was checked using Qubit (Invitrogen) and Advanced Analytical Fragment Analyzer (AATI). The SMRTbell-Polymerase Complex was prepared using Sequel™ Binding Kit 2.0 and sequenced on Sequel SMRT Cell 1M v2. A 6 h movie using the Sequel Sequencing kit 2.0. 10x Genomics libraries also constructed and sequenced on the Illumina platform to produce GemCode linked reads. All sequencing described above was outsourced to Novogene, Co., Ltd. (China). Raw reads obtained were deposited to the NCBI Short Read Archive (PRJNA587046). Sequencing depths obtained for each library are recorded in **Supplementary Table S1**.

DNA Isolation for Oxford Nanopore Sequencing

The DNA isolation was carried out in accordance to high molecular weight DNA isolation protocol as described by Mayjonade et al. (2016). Approximately, 5 g of fresh leaves were harvested from rapeseed plants at 4–6 leaf stage. This frozen leaf was immediately frozen using liquid nitrogen. This frozen leaf was subjected to mechanical grinding using a mortar and pestle. 4–5 ml of pre-heated lysis buffer [1% (w/v) PVP10, 50 nM EDTA, 1.25% (w/v) SDS, 1% (w/v) $Na_2S_2O_5$, 5 mM $C_4H_{10}O_2S_2$, 100 mM TRIS pH 8, 500 nM NaCl, 1% (v/v) Triton X-100, 1% (w/v) PVP40] to the frozen leaf samples for disrupting the cell wall. This was followed by incubation of the lysate for 30 min at 37°C. In order to precipitates sodium dodecyl sulfate (SDS) and SDS-bound proteins, 0.3 volumes of 5 M Potassium Acetate was added to the lysate and spun at 8000 *g* for 12 min at 4°C. Clean DNA was then recovered by fishing out the DNA using magnetic beads.

Size Selection and Library Preparation for Oxford Nanopore Sequencing

1–3 μg of DNA was subjected to size selection using Circulomics short-read eliminator XL kit (Circulomics, Inc., Baltimore, MD, United States) according to the manufacturer's instruction. The kit uses selective precipitation to deplete DNA fragments shorter than 40 kb. The size selected DNA was then used for the preparation of the sequencing library, using SQK-LSK109 (Oxford Nanopore Technologies) kit in accordance with the manufacturer's recommendations. Following the library preparation, DNA was finally loaded onto an Oxford Nanopore MinION flow cell (version R9.4.1) for sequencing. The raw fast5 files produced by the MinION device were then base-called using *Guppy* 3.0.3 (Oxford Nanopore Technologies) with "dna_r9.4.1_450bps_hac.cfg" model using standard parameters to generate fastq file. **Supplementary Table S2** shows the statistics of reads generated.

Optical Map Construction

DNA isolation for optical mapping was performed according to the IrysPrep™ Plant Tissue-Nuclei protocol provided by BioNano Genomics. Nearly 2 grams of young leaves were harvested from dark-treated rapeseed plants, immediately

followed by fixing with 2% formaldehyde. In order to isolate the intact nuclei, fixed leaf material was subjected to homogenization in an isolation buffer containing BME, Triton X-100 and PVP-10. Purified nuclei were then embedded into an agarose gel matrix. Finally, the DNA was recovered by melting the agarose plugs using GELase™ (Epicentre) treatment. Sequence specific nick labeling using Nt.BspQI (recognition site GCTCTTC) was performed on the isolated DNA using the IrysPrep™ Labeling-NLRS protocol by BioNano Genomics. Finally these single DNA molecules were loaded onto an IrysChip for imaging on the BioNano Genomics Irys platform. The DNA molecules were imaged using the BioNano Irys System and were computationally translated into single-molecule optical maps. Single optical maps were then assembled into a consensus map with IrysSolve pipeline (v5134) provided by BioNano Genomics, and deposited as **supplementary file** in NCBI BioProject PRJNA587046.

Genetic Maps Construction

Genetic maps were constructed for the two biparental populations Express 617 × R53 (ExR53-DH) and Express 617 × V8 (ExV8-DH) using 60K Illumina Infinium Brassica SNP array, SSR and AFLP marker data obtained from 244 and 216 lines, respectively. SNP and SNaP marker data were filtered according to Gabur et al. (2018). SSR and AFLP marker data were taken from the genetic maps produced by Radoev et al. (2008) and Basunanda et al. (2010). Genetic maps were constructed using the software *MSTMap* (Wu et al., 2008) applying the Kosambi map function. The genetic linkage map for Express 617 × SGDH14 was produced using 60K Illumina Infinium Brassica SNP array marker data obtained from 139 lines using the software package *JoinMap* 4.1 (Stam, 1993; van Ooijen, 2011) applying the Kosambi map function (Behnke et al., 2018).

Genome Assembly

To increase read accuracy, PacBio raw reads were self-aligned to generate consensus reads using *Daligner* v1.0 (Myers, 2014) using default parameters. Consensus reads were then assembled using *FALCON* (falcon-2017.11.02-16.04-py2.7) (Chin et al., 2016) to form unitigs. Unitigs were then further polished using the consensus algorithm Quiver (SMRT Link v5.0.1)[2]. Illumina short reads were used to correct small-scale errors using default parameters of *Pilon* (pilon-1.18.jar) (Walker et al., 2014). To increase contiguity, PacBio reads were used to further scaffold the unitigs (SSPACE-standard) (Boetzer et al., 2011). 10x Genomics data were first processed by trimming the first 16 bp barcode and subsequence 7 bp random primer sequence of the first mate of each pair, and then aligned to the scaffolds to form super-scaffolds using *fragScaff* (version 140324) (Adey et al., 2014). Assembly procedures described above were performed by Novogene, Co., Ltd.

The obtained super-scaffolds were corrected for large-scale chimeric regions originating from misassembly by comparing to an Express 617 BioNano optical map using *Chimericognizer* (Pan and Lonardi, 2019) where junctions of chimeric scaffolds were broken with the following parameters "-a 1.5 -b 1 -d 25 -e 50000 -h 50000 -r 80000". Scaffolds were then corrected using *Novo&Stitch* (Pan et al., 2018) with the default strict parameters that are equivalent to "-a 3000 -b 0.1 -c 10000 -d 0.5 -e 0.9 -h 25 -r 0.2".

The corrected scaffolds were then arranged into pseudochromosmes using three high-resolution SNP-based genetic maps, including Express 617 × SGDH14, Express 617 × V8 and Express 617 × R53. Weighted priority 3, 2 and 1 were given to the listed maps respectively based on the synthetic origin of parents. Pseudochromosome construction was completed using the software *ALLMAPS* (Tang et al., 2015) with the following parameters "python –m jcvi.assembly.allmaps path –mincount = 10 –links = 25". Scaffolds that map to multiple linkage groups were identified as potentially chimeric, and the breakpoints were detected using "python –m jcvi.assembly.allmaps split" and "python –m jcvi.assembly.patch refine". Corrected pseudochoromosomes were produced by repeating the *ALLMAPS* run with the broken scaffolds.

Gene Annotation

Repetitive sequences were identified using *RepeatModeler* (Smit and Hubley, 2008), a repeat family identification and modeling package which performs two *de novo* repeat-finding programs *RECON* (Bao, 2002) and *RepeatScout* (Price et al., 2005). The repeats identified were soft-masked in the assembly using *RepeatMasker* vopen-4.0.7 (Smit et al., 2013).

The gene prediction pipeline *BRAKER2* (Hoff et al., 2019) was used to train an Express 617-specific gene model and then predict genic regions. *BRAKER2* executes the gene predictor *Augustus* (Hoff and Stanke, 2013, 2018) internally. First, the proteomes of two species *Arabidopsis thaliana* (Proteome ID UP000006548) and *B. napus* (Proteome ID UP000028999) were aligned to the genome using *Genome Threader* (Gremme et al., 2005), and provided as evidence for model training in *Augustus*. The trained parameters were then used, together protein homology hints, to accurately predict genes. Fragmented predictions and potential pseudogenes were further filtered with (1) *Augustus* script "python Augustus/scripts/getAnnoFastaFromJoingenes.py –s TRUE" and (2) high identity to all 341468 proteins in the *Brassica* genus (Taxon identifier 3705) in UniProt release 2019_08 (The UniProt Consortium, 2019), with an alignment coverage of 80% to both target and query, a percentage identity of 80% and above, and -evalue 10e-5 using *BLASTP* (Camacho et al., 2009).

Evaluation of Base-Level Accuracy

To estimate the error rate of PacBio reads, two Illumina sequencing libraries (SRR10382360 and SRR10382369) were aligned to the assembly using *Bowtie2* version 2.2.6 (Langmead and Salzberg, 2012) with the following parameters "bowtie2 -I 200 -X 500 –end-to-end –no-discordant". Error rate was estimated by calculating the ratio of mismatch bases ("mismatches") to total bases of mapped reads ("bases mapped") from the output of samtools stats version 1.9 (Li et al., 2009).

Five Illumina sequencing libraries (SRR10382360, SRR10382369, SRR10382370, SRR10382371, SRR1030294) were aligned to the assembly using *Bowtie2* version 2.2.6 (Langmead and Salzberg, 2012) with the following parameters

[2]https://github.com/PacificBiosciences/GenomicConsensus

"bowtie2 –I 200 –X 500 –end-to-end –no-mixed –no-discordant". Read pairs which align perfectly were counted with the following command "cat file.sam | grep -v "^@" | cut -f1,6 | uniq -c | grep 150M | grep -vc "1 "". The same five libraries were also aligned to a subset of PacBio corrected reads.

Evaluation of Scaffolding Accuracy

Nanopore sequences (SRR10383383) were first filtered by q10 using *NanoFilt* (De Coster et al., 2018) and corrected to replace sequencing noise with consensus using *Canu* (Schmidt et al., 2017) with the following parameters "-genomeSize = 1g -correctedErrorRate = 0.105 -minReadLength = 3000 -minOverlapLength = 2000 -corOutCoverage = 200 "batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50" -ovlMerDistinct = 0.975". PacBio reads were also corrected using *Canu* (Schmidt et al., 2017) with the same parameters except for "-correctedErrorRate = 0.045". The corrected Nanopore and PacBio reads, were aligned to the assembly using *NGMLR* version 0.2.7 (Sedlazeck et al., 2018) with the following parameters "-x ont –no-smallinv".

The GOGG method (He and Bancroft, 2018), which uses the distribution pattern of gene allelic SNPs of 134 lines in a mapping population the AC *Brassica* genomics platform (He et al., 2015) to detect large structural misassemblies, was performed. The results were manually inspected for blocks with deviating patterns. The collinearity of putatively misassembled blocks with the AC pantranscriptome, *Arabidopsis thaliana* and *Thellungiella parvula* were used as supporting evidence for misjoins. Breakpoints of misjoins were resolved by inspecting the alignments of long reads with *IGV* (Robinson et al., 2011) in putatively misassembled regions. When supported by clear alteration of read coverage, the breakpoints were cut to isolate the misassembled blocks using *fastasubseq* under *Exonerate* suite (Slater and Birney, 2005). Gene annotation was updated to the corrected assembly using *flo* (Pracana et al., 2017) which implements the UCSC tool *liftover* (Kuhn et al., 2013).

Evaluation of Gene Set Completeness

The completeness of genic regions were evaluated with two standard assessment pipelines, *BUSCO* v4.0.4 (Simão et al., 2015) and *PLAZA coreGF* (van Bel et al., 2012), where the presence of highly-conserved orthologs was used to score an assembly. *BUSCO* was performed on the genome assembly using the lineage dataset brassicales_odb10. Results were plotted using the generate_plot.py script of *BUSCO*. Since the coreGF python script does not set alignment threshold, predicted proteins were first aligned to PLAZA_2.5_proteome.fasta using *BLASTP* (Camacho et al., 2009), and only alignment with above 60 percentage identity and "-evalue 10e-5 –qcov_hsp_perc 60" were used to calculate for weighted score against the "greenplants" coreGFs.

Single-ended mRNA sequencing data (SRR3134083) was aligned to the assembled genome using *HISAT2* (Kim et al., 2019) and converted to intron boundaries using *bam2hints* (Augustus 3.2.1) (Hoff and Stanke, 2013, 2018). Positions were compared using *windowBed* (v2.25.0) (Quinlan and Hall, 2010).

Predicted proteins were evaluated with presence of known protein domains using *InterproScan* v5.33-72.0 (Jones et al., 2014) with the following parameters "interproscan.sh -appl TIGRFAM, PANTHER, Pfam, PrositeProfiles, PrositePatterns –iprlookup –goterms –pa".

The predicted coding sequences were also aligned to the A and C genome-based ordered pan-transcriptome (He et al., 2015) using *BLASTN* (Camacho et al., 2009) with the following parameters "-qcov_hsp_perc 60, -evalue 10e-5" and only alignments coverage of 60% to both target and query, a percentage identity of 60% and above were counted.

Genome-to-Genome Comparison

Whole genome alignment between assemblies was performed using the graphical interface of *D-Genies* (Cabanettes and Klopp, 2018), which invokes *Minimap2* (Heng Li, 2018) internally and generates dotplots. Dotplots were displayed by applying a match size filtering, where matches that were grouped in the bins of smaller sizes (first and second out of 7 bins) were removed. Genomes used were *B. napus* Darmor-*bzh* v 4.1 (Chalhoub et al., 2014), *B. rapa* and *B. oleracea* (Belser et al., 2018).

Collinearity between genomes was identified by first obtaining orthologous genes and then these genes were used as anchor to detect synteny and collinearity using *MCScanX* (Wang et al., 2012). Since the *MCScanX* recommends around five hits per transcript, the alignment run was performed in two steps: (1) aligning all transcripts of both genomes to themselves and to each other using *BLASTN* (Camacho et al., 2009) with following parameters (2) filtering hits by alignment coverage and percentage identity of 80% and above. *MCScanX* was performed with parameters "match_score 50, match_size 10, gap_penalty -1, overlap_window 10, e_value 1e-05, max_gaps 10". The output was then plotted for visualization using *SynVisio*[3] to display blocks with final score of 10000 and above.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI BioProject PRJNA587046 (sequencing reads and optical map) and Zenodo https://doi.org/10.5281/zenodo.3524259 (unpublished genetic maps, assembled genome and annotation results).

AUTHOR CONTRIBUTIONS

HL conducted the analysis and drafted the manuscript. HC and CO constructed the optical map and HC sequenced the Nanopore long reads. CO, FD, and AA constructed genetic maps. RS and HL conceived the study. All critically reviewed and edited the manuscript.

[3]https://kiranbandi.github.io/synvisio/

FUNDING

The work described was initiated within the framework project IRFFA: Improved Rapeseed as Fish Feed in Aquaculture. Funding was provided by grant 031B0357A-D from the German Federal Ministry of Education and Research (BMBF).

ACKNOWLEDGMENTS

The authors acknowledge Ian Bancroft and Zhesi He from the Department of Biology in the University of York (Heslington, York, United Kingdom) for assessing the genome assembly with the GOGG pipeline, and Stavros Tzigos and Andreas Welke (Justus Liebig University Giessen) for technical assistance in the laboratory and greenhouse. The authors acknowledge the BBSRC BRAVO project (UK) as source of part of the RNAseq data use.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00496/full#supplementary-material

REFERENCES

Adey, A., Kitzman, J. O., Burton, J. N., Daza, R., Kumar, A., Christiansen, L., et al. (2014). In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* 24, 2041–2049. doi: 10.1101/gr.178319.114

Allender, C. J., and King, G. J. (2010). Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers. *BMC Plant Biol.* 10:54. doi: 10.1186/1471-2229-10-54

Appelqvist, L.-A., and Ohlson, R. (1972). *Rapeseed : Cultivation, Composition, Processing and Utilization*. Amsterdam: Elsevier Publishing Company.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97. doi: 10.1126/science.aan0032

Badani, A. G., Snowdon, R. J., Wittkop, B., Lipsa, F. D., Baetzel, R., Horn, R., et al. (2006). "Colocalization of a partially dominant gene for yellow seed colour with a major QTL influencing acid detergent fibre (ADF) content in different crosses of oilseed rape (*Brassica napus*). *Genome* 49, 1499–1509. doi: 10.1139/g06-091

Bao, Z. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi: 10.1101/gr.88502

Basunanda, P., Radoev, M., Ecke, W., Friedt, W., Becker, H. C., and Snowdon, R. J. (2010). Comparative mapping of quantitative trait loci involved in heterosis for seedling and yield traits in oilseed rape (*Brassica napus* L.). *Theor. Appl. Genet.* 120, 271–281. doi: 10.1007/s00122-009-1133-z

Bayer, P. E., Edwards, D., and Batley, J. (2018). Bias in resistance gene prediction due to repeat masking. *Nat. Plants* 4, 762–765. doi: 10.1038/s41477-018-0264-0

Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C. K., Yuan, Y., Lee, H., et al. (2017). Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnol. J.* 15, 1602–1610. doi: 10.1111/pbi.12742

Becker, H. C., Engqvist, G. M., and Karlsson, B. (1995). Comparison of rapeseed cultivars and resynthesized lines based on allozyme and RFLP markers. *Theor. Appl. Genet.* 91, 62–67. doi: 10.1007/BF00220859

Behnke, N., Suprianto, E., and Möllers, C. (2018). A major QTL on chromosome C05 significantly reduces acid detergent lignin (ADL) content and increases seed oil and protein content in oilseed rape (*Brassica napus* L.). *Theor. Appl. Genet.* 131, 2477–2492. doi: 10.1007/s00122-018-3167-6

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887. doi: 10.1038/s41477-018-0289-4

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? *GigaScience* 8:giz068. doi: 10.1093/gigascience/giz068

Bhat, R. S., Shirasawa, K., Monden, Y., Yamashita, H., and Tahara, M. (2020). "Developing transposable element marker system for molecular breeding," in *Legume Genomics*, eds M. Jain, and R. Garg, (New York, NY: Springer), 233–251. doi: 10.1007/978-1-0716-0235-5_11

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

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

Bus, A., Körber, N., Snowdon, R. J., and Stich, B. (2011). Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. *Theor. Appl. Genet.* 123, 1413–1423. doi: 10.1007/s00122-011-1676-7

Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6:e4958. doi: 10.7717/peerj.4958

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Chalhoub, B., Denoeud, F., Liu, S. I, Parkin, A. P., 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

Chelaifa, H., Monnier, A., and Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina* × *Townsendii* and *Spartina Anglica* (Poaceae). *New Phytol.* 186, 161–174. doi: 10.1111/j.1469-8137.2010.03179.x

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

Coate, J. E., Bar, H., and Doyle, J. J. (2014). Extensive translational regulation of gene expression in an allopolyploid (*Glycine dolichocarpa*). *Plant Cell* 26, 136–150. doi: 10.1105/tpc.113.119966

Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H., and Stein, L. (2005). Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.* 21, 673–682. doi: 10.1016/j.tig.2005.09.009

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

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

Delourme, R., Laperche, A., Bouchet, A.-S., Jubault, M., Paillard, S., Manzanares-Dauleux, M.-J., et al. (2018). "Genes and quantitative trait loci mapping for major agronomic traits in *Brassica napus* L," in *The Brassica Napus Genome*, eds S. Liu, R. Snowdon, and B. Chalhoub, (Cham: Springer International Publishing), 41–85. doi: 10.1007/978-3-319-43694-4_3

Diers, B. W., and Osborn, T. C. (1994). Genetic diversity of oilseed *Brassica napus* germ plasm based on restriction fragment length polymorphisms. *Theor. Appl. Genet.* 88, 662–668. doi: 10.1007/BF01253968

Edger, P. P., McKain, M. R., Bird, K. A., and van Buren, R. (2018). Subgenome assignment in allopolyploids: challenges and future directions. *Curr. Opin. Plant Biol.* 42, 76–80. doi: 10.1016/j.pbi.2018.03.006

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

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

Food and Agriculture Organization of the United Nations (2019). *FAOSTAT 2017*. Available online at: http://faostat.fao.org (accessed September, 2019).

Friedt, W., and Snowdon, R. (2009). "Oilseed rape," in *Oil Crops*, eds J. Vollmann, and I. Rajcan, (New York, NY: Springer), 91–126. doi: 10.1007/978-0-387-77594-4_4

Friedt, W., Tu, J., and Fu, T. (2018). "Academic and economic importance of *Brassica napus* rapeseed," in *The Brassica Napus Genome*, eds S. Liu, R. Snowdon, and B. Chalhoub, (Cham: Springer International Publishing), 1–20. doi: 10.1007/978-3-319-43694-4_1

Gabur, I., Chawla, H. S., Liu, X., Kumar, V., Faure, S., von Tiedemann, A., et al. (2018). Finding invisible quantitative trait loci with missing data. *Plant Biotechnol. J.* 16, 2102–2112. doi: 10.1111/pbi.12942

Gabur, I., Chawla, H. S., Snowdon, R. J., and Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.* 132, 733–750. doi: 10.1007/s00122-018-3233-0

Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E., and Osborn, T. C. (2007). Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* 19, 3403–3417. doi: 10.1105/tpc.107.054346

Gao, C. (2018). The future of CRISPR technologies in agriculture. *Nat. Rev. Mol. Cell Biol.* 19, 275–276. doi: 10.1038/nrm.2018.2

Gazave, E., Tassone, E. E., Ilut, D. C., Wingerson, M., Datema, E., Witsenboer, H. M., et al. (2016). Population genomic analysis reveals differential evolutionary histories and patterns of diversity across subgenomes and subpopulations of *Brassica napus* L. *Front. Plant Sci.* 7:525. doi: 10.3389/fpls.2016.00525

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

Gremme, G., Brendel, V., Sparks, M. E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *Inform. Softw. Technol.* 47, 965–978. doi: 10.1016/j.infsof.2005.09.005

Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., et al. (2012). Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nat. Biotechnol.* 30, 798–802. doi: 10.1038/nbt.2302

He, Z., and Bancroft, I. (2018). Organization of the genome sequence of the polyploid crop species *Brassica juncea*. *Nat. Genet.* 50, 1496–1497. doi: 10.1038/s41588-018-0239-0

He, Z., Cheng, F., Li, Y., Wang, X., Parkin, I. A., Chalhoub, B., et al. (2015). Construction of *Brassica* A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. *Data Brief.* 4, 357–362. doi: 10.1016/j.dib.2015.06.016

He, Z., Wang, L., Harper, A., Havlickova, L., Pradhan, A., Parkin, I., and Bancroft, I. (2017). Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* 15, 594–604. doi: 10.1111/pbi.12657

Hoff, K. J., and Stanke, M. (2013). WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* 41, W123–W128. doi: 10.1093/nar/gkt418

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-genome annotation with BRAKER," in *Gene Prediction*, ed. M. Kollmar, (New York, NY: Springer), 65–95. doi: 10.1007/978-1-4939-9173-0_5

Hoff, K. J., and Stanke, M. (2018). Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinform.* 65:e57. doi: 10.1002/cpbi.57

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

Jiao, W. B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., et al. (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* 27, 778–786. doi: 10.1101/gr.213652.116

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., et al. (2005). Evolution of genome size in brassicaceae. *Ann. Bot.* 95, 229–235. doi: 10.1093/aob/mci016

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kim, S., Park, J., Yeom, S. I., Kim, Y. M., Seo, E., Kim, K. T., et al. (2017). New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol.* 18:210. doi: 10.1186/s13059-017-1341-9

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36, 1174–1182. doi: 10.1038/nbt.4277

Korlach, J. (2013). *Understanding Accuracy in SMRT$Sequencing: Pacific Biosciences*. Available online at: https://www.mscience.com.au/upload/pages/pacbioaccuracy/perspective-understanding-accuracy-in-smrt-sequencing.pdf

Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief. Bioinform.* 14, 144–161. doi: 10.1093/bib/bbs038

Lai, X., Schnable, J. C., Liao, Z., Xu, J., Zhang, G., Li, C., et al. (2017). Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. *BMC Genomics* 18:702. doi: 10.1186/s12864-017-4103-x

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

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, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, X., Guo, K., Zhu, X., Chen, P., Li, Y., Xie, G., et al. (2017). Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. *BMC Genomics* 18:55. doi: 10.1186/s12864-016-3454-z

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., et al. (2018). Genome sequence of the progenitor of wheat a subgenome *Triticum urartu*. *Nature* 557, 424–428. doi: 10.1038/s41586-018-0108-0

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, A. P. I., et al. (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* 5:3930. doi: 10.1038/ncomms4930

Lloyd, A., Blary, A., Charif, D., Charpentier, C., Tran, J., Balzergue, S., et al. (2018). Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. *New Phytol.* 217, 367–377. doi: 10.1111/nph.14836

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444

Lombard, V., and Delourme, R. (2001). A consensus linkage map for rapeseed (*Brassica napus* L.): construction and integration of three individual maps from DH populations. *Theor. Appl. Genet.* 103, 491–507. doi: 10.1007/s001220100560

Lukens, L. N., Pires, J. C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus* allopolyploids. *Plant Physiol.* 140, 336–348. doi: 10.1104/pp.105.066308

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

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

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., and Lysak, M. A. (2010). Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* 22, 2277–2290. doi: 10.1105/tpc.110.074526

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., et al. (2016). Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *BioTechniques* 61, 203–205. doi: 10.2144/000114460

Myers, G. (2014). "Efficient local alignment discovery amongst noisy long reads," in *Algorithms in Bioinformatics*, eds D. Brown, and B. Morgenstern, (Berlin: Springer), doi: 10.1007/978-3-662-44753-6_5

Nagaharu, U. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot.* 7, 389–452.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. *Nat. Rev. Genet.* 14, 157–167. doi: 10.1038/nrg3367

Obermeier, C., Hossain, M. A., Snowdon, R., Knüfer, J., von Tiedemann, A., and Friedt, W. (2013). Genetic analysis of phenylpropanoid metabolites associated with resistance against *Verticillium longisporum* in *Brassica napus*. *Mol. Breed.* 31, 347–361. doi: 10.1007/s11032-012-9794-8

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

Pan, W., and Lonardi, S. (2019). Accurate detection of chimeric contigs via bionano optical maps. *Bioinformatics* 35, 1760–1762. doi: 10.1093/bioinformatics/bty850

Pan, W., Wanamaker, S. I., Ah-Fong, A. M. V., Judelson, H. S., and Lonardi, S. (2018). Novo&Stitch: accurate reconciliation of genome assemblies via optical maps. *Bioinformatics* 34, i43–i51. doi: 10.1093/bioinformatics/bty255

Parkin, I. A., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C., et al. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* 171, 765–781. doi: 10.1534/genetics.105.042093

Piquemal, J., Cinquin, E., Couton, F., Rondeau, C., Seignoret, E., Doucet, I., et al. (2005). Construction of an oilseed rape (*Brassica napus* L.) genetic map with SSR markers. *Theor. Appl. Genet.* 111, 1514–1523. doi: 10.1007/s00122-005-0080-6

Pires, J. C., Zhao, J., Schranz, M. E., Leon, E. J., Quijada, P. A., Lukens, L. N., et al. (2004). Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biol. J. Linn. Soc.* 82, 675–688. doi: 10.1111/j.1095-8312.2004.00350.x

Pracana, R., Priyam, A., Levantis, I., Nichols, R. A., and Wurm, Y. (2017). The fire ant social chromosome supergene variant SB shows low diversity but high divergence from SB. *Mol. Ecol.* 26, 2864–2879. doi: 10.1111/mec.14054

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl. 1), i351–i358. doi: 10.1093/bioinformatics/bti1018

Qian, L., Voss-Fels, K., Cui, Y., Jan, H. U., Samans, B., Obermeier, C., et al. (2016). Deletion of a stay-green gene associates with adaptive selection in *Brassica napus*. *Mol. Plant* 9, 1559–1569. doi: 10.1016/j.molp.2016.10.017

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

Radoev, M., Becker, H. C., and Ecke, W. (2008). Genetic analysis of heterosis for yield and yield components in rapeseed (*Brassica napus* L.) by quantitative trait locus mapping. *Genetics* 179, 1547–1558. doi: 10.1534/genetics.108.089680

Renny-Byfield, S., Gong, L., Gallagher, J. P., and Wendel, J. F. (2015). Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol. Biol. Evol.* 32, 1063–1071. doi: 10.1093/molbev/msv001

Rigal, M., Becker, C., Pélissier, T., Pogorelcnik, R., Devos, J., Ikeda, Y., et al. (2016). Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in *Arabidopsis thaliana* F1 epihybrids. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2083–E2092. doi: 10.1073/pnas.1600672113

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Rousseau-Gueutin, M., Morice, J., Coriton, O., Huteau, V., Trotoux, G., Nègre, S., et al. (2017). The impact of open pollination on the structural evolutionary dynamics, meiotic behavior, and fertility of resynthesized allotetraploid *Brassica napus* L. *G3 (Bethesda)* 7, 705–717. doi: 10.1534/g3.116.036517

Salmon, A., Ainouche, M. L., and Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Mol. Ecol.* 14, 1163–1175. doi: 10.1111/j.1365-294X.2005.02488.x

Samans, B., Chalhoub, B., and Snowdon, R. J. (2017). Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2017.02.0013

Schiessl, S., Huettel, B., Kuehn, D., Reinhardt, R., and Snowdon, R. (2017). Post-polyploidisation morphotype diversification associates with gene copy number variation. *Sci. Rep.* 7:41845. doi: 10.1038/srep41845

Schmidt, M. H., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., et al. (2017). De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* 29, 2336–2348. doi: 10.1105/tpc.17.00521

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

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15, 461–468. doi: 10.1038/s41592-018-0001-7

Sharpe, A. G. I, Parkin, A. P., Keith, D. J., and Lydiate, D. J. (1995). Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). *Genome* 38, 1112–1121. doi: 10.1139/g95-148

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Slater, G., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6:31. doi: 10.1186/1471-2105-6-31

Smit, A. F. A., and Hubley, R. (2008). *RepeatModeler Open-1.0*. http://www.repeatmasker.org (accessed October, 2018).

Smit, A. F. A., Hubley, R., and Green, P. (2013). *RepeatMasker Open-4.0*. http://www.repeatmasker.org (accessed October, 2019).

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

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S. V., Obermeier, C., et al. (2017). Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. *Plant Biotechnol. J.* 15, 1478–1489. doi: 10.1111/pbi.12732

Stein, A., Wittkop, B., Liu, L., Obermeier, C., Friedt, W., and Snowdon, R. J. (2013). "Dissection of a major QTL for seed colour and fibre content in *Brassica napus* reveals colocalization with candidate genes for phenylpropanoid biosynthesis and flavonoid deposition. *Plant Breed.* 132, 382–389. doi: 10.1111/pbr.12073

Sun, F., Fan, G., Hu, Q., Zhou, Y., Guan, M., Tong, C., et al. (2017). "The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. *Plant J.* 92, 452–468. doi: 10.1111/tpj.13669

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., et al. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 16:3. doi: 10.1186/s13059-014-0573-1

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant* 12, 156–169. doi: 10.1016/j.molp.2018.12.016

The International Wheat Genome Sequencing Consortium (IWGSC), Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

UniProt Consortium, (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Udall, J. A., Quijada, P. A., and Osborn, T. C. (2005). Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L. *Genetics* 169, 967–979. doi: 10.1534/genetics.104.033209

van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., et al. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* 158, 590–600. doi: 10.1104/pp.111.189514

van Ooijen, J. W. (2011). Multipoint maximum likelihood mapping in a full-Sib family of an outbreeding species. *Genet. Res. (Camb)* 93, 343–349. doi: 10.1017/S0016672311000279

Veeckman, E., Ruttink, T., and Vandepoele, K. (2016). Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* 28, 1759–1768. doi: 10.1105/tpc.16.00349

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection

and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 37, 124–126. doi: 10.1038/s41587-018-0004-z

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies: fig. 1. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383

Wu, Y., Bhat, P. R., Close, T. J., and Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* 4:e1000212. doi: 10.1371/journal.pgen.1000212

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

Zhao, G., Zou, C., Li, K., Wang, K., Li, T., Gao, L., et al. (2017). The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nat. Plants* 3, 946–955. doi: 10.1038/s41477-017-0067-8

Conflict of Interest: AA and FD were employed by the company NPZ Innovation GmbH.

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.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 June 2020
doi: 10.3389/fpls.2020.00701

BRIDGE – A Visual Analytics Web Tool for Barley Genebank Genomics

Patrick König[1], Sebastian Beier[1], Martin Basterrechea[1], Danuta Schüler[1], Daniel Arend[1], Martin Mascher[1,2], Nils Stein[1,3], Uwe Scholz[1]* and Matthias Lange[1]*

[1] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [2] German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany, [3] Center for Integrated Breeding Research, Georg-August University, Göttingen, Germany

Genebanks harbor a large treasure trove of untapped plant genetic diversity. A growing world population and a changing climate require an increase in the production and development of stress resistant plant cultivars while decreasing the acreage. These requirements for improved plant cultivars can be supported by the broader exploitation of plant genetic resources (PGR) as inputs for genomics-assisted breeding. To support this process we have developed BRIDGE, a data warehouse and exploratory data analysis tool for genebank genomics of barley (*Hordeum vulgare* L.). Using efficient technologies for data storage, data transfer and web development, we facilitate access to digital genebank resources of barley by prioritizing the interactive and visual analysis of integrated genotypic and phenotypic data. The underlying data resulted from a barley genebank genomics study cataloging sequence and morphological data of 22,626 barley accessions, mainly from the German Federal *ex situ* genebank. BRIDGE consists of interactively coupled modules to visualize integrated, curated and quality checked data, such as variation data, results of dimensionality reduction and genome wide association studies (GWAS), phenotyping results, passport data as well as the geographic distribution of germplasm samples. The core component is a manager for custom collections of germplasm. A search module to find and select germplasm by passport and phenotypic attributes is included as well as modules to export genotypic data in gzip-compressed variant call format (VCF) files and phenotypic data in MIAPPE-compliant ISA-Tab files. BRIDGE is accessible at the following URL: https://bridge.ipk-gatersleben.de.

Keywords: barley, plant genetic resources, genebank genomics, visual analytics, data visualization, phenotyping, genotyping, data warehouse

OPEN ACCESS

Edited by:
Ingo Ebersberger, Goethe-Universität Frankfurt am Main, Germany

Reviewed by:
Stefan Simm, Goethe Business School, Germany
Xiaoli Jin, Zhejiang University, China

****Correspondence:***
Patrick König
koenig@ipk-gatersleben.de
Uwe Scholz
scholz@ipk-gatersleben.de

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *30 January 2020*
Accepted: *04 May 2020*
Published: *11 June 2020*

Citation:
König P, Beier S, Basterrechea M, Schüler D, Arend D, Mascher M, Stein N, Scholz U and Lange M (2020) BRIDGE – A Visual Analytics Web Tool for Barley Genebank Genomics. Front. Plant Sci. 11:701. doi: 10.3389/fpls.2020.00701

INTRODUCTION

Cereal grasses like barley, rye and wheat are the main nutrition source for human calorie intake in the world (FAO, 2018). With its diploid genome, inbreeding feature and in comparison to its relatives wheat and rye rather small genome size of 5.1 Gbp, barley is an excellent model for basic and applied research in the *Triticeae* tribe. Large collections of plant genetic resources for food and agriculture (PGRFA) of diverse barley genotypes and phenotypes have been described in the literature (Ullrich, 2010). The German Federal *ex situ* genebank hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research in Gatersleben hosts more than 22,000 barley accessions

consisting of wild relatives, landraces, and breeding material collected over the past 70 years (Nagel et al., 2009; González et al., 2018; Milner et al., 2019). In prospect of the upcoming challenges of continued growth of the world population, climate change, and increasing scarcity of resources like arable land, water, and fertilizers enhances the pressure on agriculture to provide humankind with sufficient food (Pachauri et al., 2015; FAO, 2018). PGRFA hold a promise for the way of responding to this pressure through crop improvement and yield increase per hectare. The necessary continued crop improvement can be achieved by modern plant breeding methods like marker assisted selection (Varshney et al., 2005; Crossa et al., 2017), reverse breeding (Dirks et al., 2009), and genome engineering/editing (Voytas and Gao, 2014). These methods allow to increasingly benefit from the putative advantageous alleles of the PGRFA, which have not yet been used in recent breeding efforts. To accompany this process, information systems are needed that integrate the extensive amounts of derived multi-omics data and make them accessible and available in a FAIR manner (Wilkinson et al., 2016). Based on the integrated datasets of these information systems, decisions for the optimization of the breeding process and curation of passport data can be derived.

Due to constantly decreasing costs for DNA sequencing, ever more extensive sequencing projects can be carried out at constant or even further decreasing costs (Shendure and Ji, 2008; Shendure et al., 2017). This leads to constantly increasing amounts of raw, processed, and analyzed data that has to be stored, curated and integrated (Muir et al., 2016). Advancement in the field of sequencing have led to different experimental setups like resequencing on a population scale with various complexity reduction methods to derive molecular fingerprints of whole breeding panels (Varshney et al., 2009; Poland and Rife, 2012; Jarquín et al., 2014). Furthermore, even more ambitious projects have been initiated to generate many different reference genome assemblies to fully describe the pangenome (Weigel and Mott, 2009; Hirsch et al., 2014; Vernikos et al., 2015).

The field of non-invasive plant imaging has also seen tremendous advancements, with fully automated imaging systems able to produce multi- and hyperspectral images (Fiorani and Schurr, 2013; Coppens et al., 2017; Pieruschka and Schurr, 2019). However, the produced data for phenotyping experiments can in general be more heterogeneous than for genotyping. Expressed phenotypes can vary not only between genotypes but also between cell types (Houle et al., 2010). In addition, because of the high plasticity of plants, where different environments can trigger a single genotype to express different phenotypes, it is imperative to collect a multitude of heterogeneous data (Bolger et al., 2019).

There are well established and community accepted repositories, data deposition, and data exchange formats for genomic data, such as the European Nucleotide Archive (ENA) (Silvester et al., 2018), the National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators, 2017), and the DNA Data Bank of Japan (DDBJ) (Tateno, 2002), which are part of the International Nucleotide Sequence Database Collaboration (Cochrane et al., 2016). In contrast, there is yet no comparable central, dedicated database for phenotypic data, especially for plants. Rather, there are a large number of repositories with different specific focuses and data formats (Coppens et al., 2017). To supersede many different and incompatible text based formats for data storage and data exchange an international committee has agreed upon recommendations for a set of minimal information about plant phenotyping experiments (MIAPPE) (Krajewski et al., 2015; Ćwiek-Kupczyńska et al., 2016). An established way to digitize the results of biological and life science experiments for storage is the ISA framework (Sansone et al., 2012) and in particular the "ISA-Tab" file format (Rocca-Serra et al., 2010). In the plant science community, there is a common effort to establish the usage of MIAPPE-compliant ISA-Tab files for the export and transfer of data for plant phenotyping experiments (Krajewski et al., 2015).

The combination of genomic and phenotypic characteristics form a high-dimensional data space. Traditional genebank web portals and germplasm data warehouses, like e.g., GBIS (Oppermann et al., 2015) and EURISCO (Weise et al., 2017), have so far not focused on the visualization and explorative data analysis of integrated, high-dimensional genotypic and phenotypic datasets of entire germplasm collections. This is also the case because the specific functional scope of such genebank portals has historically been developed primarily with regard to basic genebank management tasks like seed management and reproduction of PGRs, but is not geared toward the integration or interoperability to multi-omics data repositories, visual analytics or exploratory data analysis.

Here, we present BRIDGE, a web application built to help explore and analyze the results of research data from a comprehensive barley geno- and phenotyping project of a panel of 22,626 barley (*Hordeum vulgare* L.) accessions. Genotypic data were determined for 22,621 accessions and stored in a relational database management system (RDBMS). Accessing this data source using the web frontend, the user is able to navigate between phenotypic traits, genetic and genomic information from genotyping-by-sequencing derived single nucleotide polymorphism (SNP) profiles, and can correlate them with passport information in several easy-to-understand graphical visualizations of the data. Furthermore, BRIDGE is directly connected to the information management system of the German Federal *ex situ* genebank hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, Germany. Users are able to order seed material based on their custom germplasm selections.

MATERIALS AND METHODS

Application Architecture

BRIDGE follows the client-server architecture model. The client is built as a web frontend (**Figure 1**) based on the concept of single-page applications (SPA) utilizing HTML, CSS, JavaScript, and established web technology standards like a RESTful API for communication with the server-side part of the application. The web frontend is implemented in a model-view-controller (MVC) pattern (Leff and Rayfield, 2001) based modular architecture.

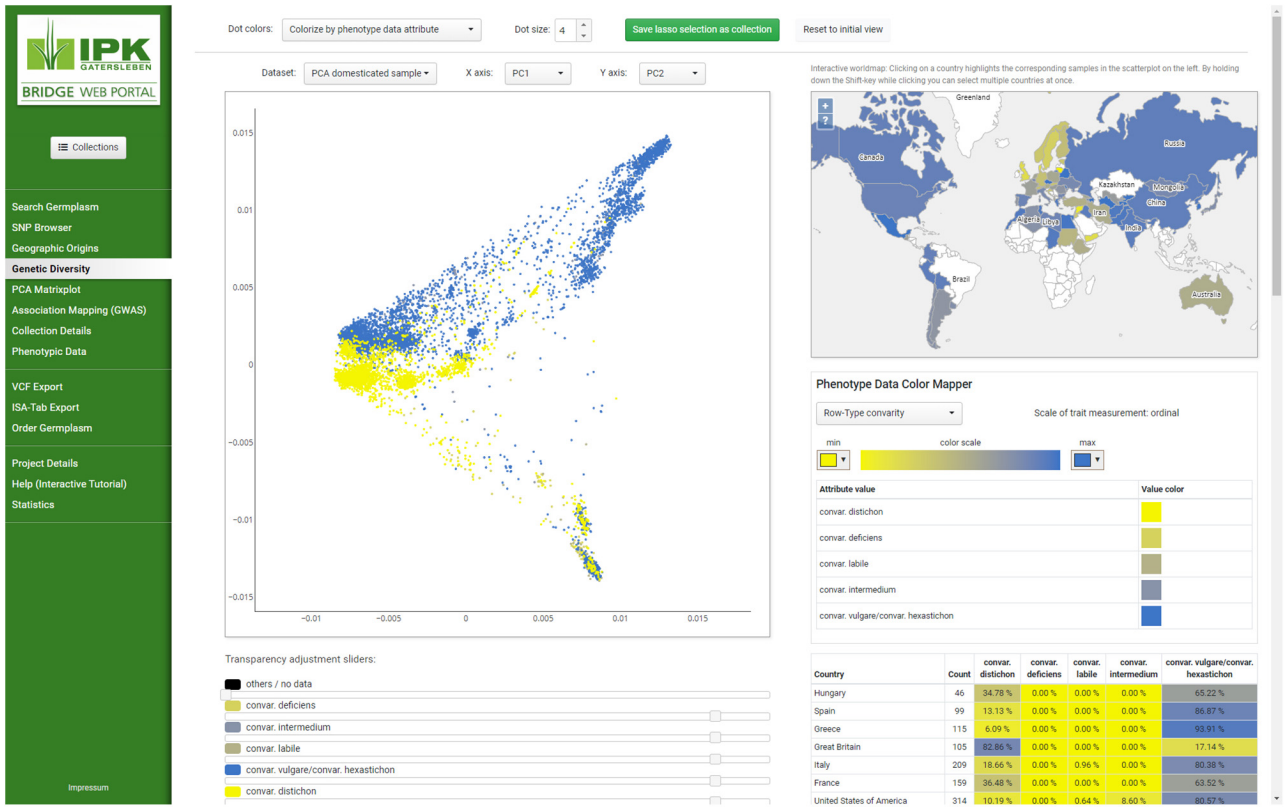

FIGURE 1 | Screenshot of the layout and BRIDGE user interface. Navigating to different parts of the page can be done by selecting the menu items on the left which are clustered into thematic groups and separated by horizontal lines to give users a visual structure. Changing from one menu item to the next does not change the appearance of the web portal, whereby both the layout remains intact and the applied settings are transferred to the entire page. This behavior is compliant to the single-page application (SPA) paradigm. As example "Genetic Diversity" from the navigation menu has been selected to present the colorization of subsets in both PCA plot and the distribution of accessions on a linked world map. Parameters of the visualization can be changed with the buttons, sliders and select boxes throughout the page. Detailed information regarding the selected subsets can be found underneath the plots.

The graphical user interface uses jQuery in version 3.2.1 (De Volder, 2006) and Vue.js in version 2.6.10 (Filipova, 2016) for the rendering of user interface widgets, the interaction with the Document Object Model (DOM) and the handling of events like user inputs. The interactive scatterplots are implemented with the Plotly.js library in version 1.43.2 (Sievert et al., 2017), which uses WebGL for hardware-accelerated high-performance rendering of plots with millions of data points. The world map was implemented with OpenLayers in version 5.3.0 (Hazzard, 2011). The client consists of a core library, which is extended by additional loosely coupled modules for each visualization or data export feature. The core library provides baseline functionality like the management of sample collections across the different modules. Due to the usage of advanced web technology standards such as HTML5 and CSS3 in the frontend, access to the webportal requires recent versions of web browsers such as Mozilla Firefox, Google Chrome, Apple Safari or Microsoft Edge. The server-side backend of BRIDGE was built as a Java Virtual Machine application using the Grails web application framework in version 3.3.8 (Smith and Ledbrook, 2009). The architecture of the BRIDGE system is shown in **Figure 2**.

Basic Usability Concepts

BRIDGE uses a concept of "named collections" of germplasm samples as the foundation for an interactive and exploratory data analysis workflow. A named collection is a user-defined, private and re-usable set of germplasm samples that can feed into the modules for visual analytics and data export. Each collection can be assigned an individual color, which is later used in the data visualization as a tagging color. Collections can be compiled by saving the result of a germplasm search (**Figure 3A**) or by saving the result of a lasso selection in a scatterplot (**Figure 3C**) or on the world map (**Figure 3D**). The user is notified if a new collection intersects with a previously saved collection (**Figure 4D**) and can decide whether to include or exclude the shared samples. It is also possible to save only the intersection itself while omitting all other samples, allowing the iterative refinement of collections.

Another concept used in BRIDGE is the synchronization of visualization modules called "interactive brushing and linking," a technique that combines different visualization methods to achieve a greater benefit compared to the standalone usage of single visualization methods (Keim, 2002). Basically, a change of parameters in one visualization directly influences the visualization of data in other visualizations. In this context,

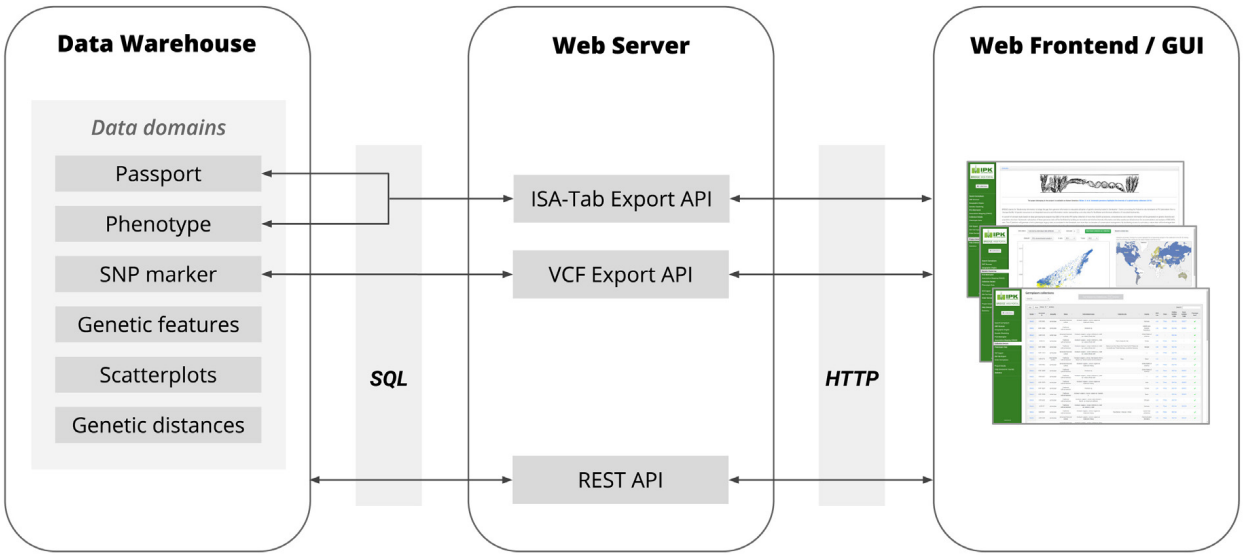

FIGURE 2 | The basic architecture of BRIDGE showing the flow of data between the Data Warehouse, Web Server and Web Frontend with all available data domains. The Data Warehouse consists of several SQL tables in the in-house ORACLE RDBMS. The communication between the Data Warehouse and Web Server is based on the JDBC interface using SQL as query language. The Web Server provided a REST API for the Web Frontend to deliver the data points for all data visualization modules and to respond to requests for search queries initiated by a user via the Web Frontend. Furthermore, the Web Server provides APIs for the export of data of genetic variants as VCF files and the export of phenotypic data in MIAPPE-compliant ISA-Tab archives. The communication between the Web Server and Web Frontend is based on the HTTP protocol.

brushing is understood as highlighting the same subset of data in dynamically coupled visualizations. This functionality is implemented by an automatic color tagging of data points of the user-defined germplasm sample collections in the various visualization modules (**Figure 4**).

Available Data

The provided data in BRIDGE are the results of the study published by Milner et al. (2019) who compiled molecular passport data for the entire barley germplasm collection of the IPK genebank and several hundred additional accessions from genebanks that were incorporated in the corresponding research project. The available data in BRIDGE consists of passport data of 22,626 germplasm samples, phenotypic data of 9,527 germplasm samples, SNP matrices, visualization of the genetic diversity space by PCA and t-SNE (Maaten and van der Hinton, 2008) and Manhattan plots of GWAS results. As a result of ongoing research and maintenance, passport data are continuously curated by the genebanks and can therefore vary over time. To ensure consistency with the source data of the study of Milner et al. (2019), the passport data used in BRIDGE is a snapshot taken at the beginning of the study. A categorized summary of the data stock is listed in **Table 1**. The available attributes and completeness of passport records is listed in **Table 2**. Additional attributes are the affiliation of a germplasm sample to one of the three Core sets, the accession number prefix and the row type derived from the full botanical name of the accession. This row type attribute is present in almost all germplasm samples, but may contain errors due to data problems in the full botanical names of historical passport data. It is important to note that this row type, which is derived from passport data, does not have to be the same as the phenotyped row type (for example because of intra-accession heterogeneity or a mislabeling of the subspecies attribute). A detailed tabular summary of all available phenotypic traits with all possible values and the count of germplasm samples for each distinct trait value is listed in **Table 3**. We note that the number of observed phenotypes and passport information is very unbalanced. This is due to the limited project resources in terms of available working time for manual phenotyping.

Genetic variation data are present in the form of two SNP matrices: one with unimputed data of 187,713 SNPs and 22,621 genotypes and another with imputed data of 306,049 SNPs and 20,458 genotypes. SNP positions are given relative to the first version reference sequence assembly of cv. Morex (Mascher et al., 2017). As data for the genome annotation in the SNP browser the data sets of structural and functional information of the International Barley Sequencing Consortium was used (IBSC, 2016a,b).

Results of GWAS are available for the following analyzed traits:

- Grain hull (covered vs. naked),
- Row-type (two-rowed vs. six-rowed),
- Awn roughness,
- Resistance to Barley Mild Mosaic Virus (BaMMV),
- Flowering time.

The dimensionality reduction of the SNP data with the t-SNE algorithm was calculated using the function *TSNE()* of the *scikit-learn* package (Pedregosa et al., 2011) for Python. Except for

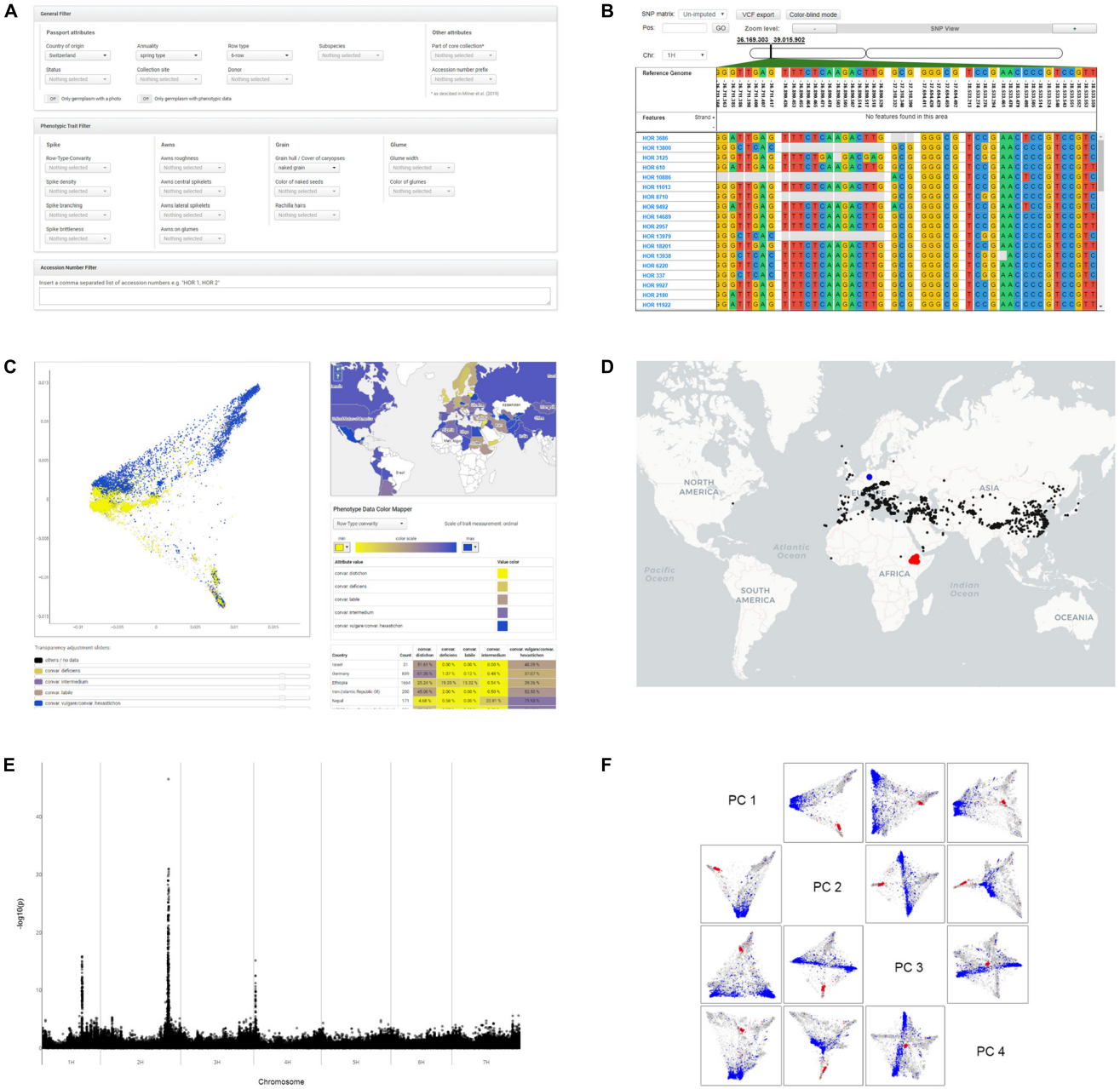

FIGURE 3 | Screenshots as illustrations of the most important search and visualization features, which can be used as an entry point for exploratory data analysis. **(A)** Data Filters: In the "Search Germplasm" panel, the user can search for germplasm by filtering passport attributes, phenotypic traits and SNP markers. **(B)** Genomic Diversity Visualization: In the integrated SNP browser users can inspect and subset the SNP matrices visually for different collections of germplasm. **(C)** Combined Interactive Visualizations: This tool enables the correlation of results of dimensionality reduction algorithms like PCA or t-SNE on the SNP data with countries of origin and phenotypic traits of the germplasm. **(D)** Interactive World Map: This tool allows the user to create lasso selections of geo-localized germplasm or to highlight user-defined germplasm collections with their specific tagging colors. **(E)** Manhattan Plots: This tool provides interactive plots of GWAS analysis results where each SNP data point is linked to the SNP browser. The user can click on a SNP data point and is then automatically guided to the corresponding genomic location in the SNP browser. **(F)** PCA Scatterplot Matrix: This visualization tool allows visual inspection of the first four principal components while highlighting user-defined germplasm collections with their specific tagging colors. It also allows to save custom lasso-selection of data points as a "named collection" of germplasm.

the parameter "perplexity" the function was called with default values for all parameters. The parameter "perplexity" was varied to shift between local and global structure preservation. To feed the *TSNE()* function, the VCF file was converted into a NumPy array (van der Walt et al., 2011) using the function *read_vcf()* of the *scikit-allel* package (Miles et al., 2019).

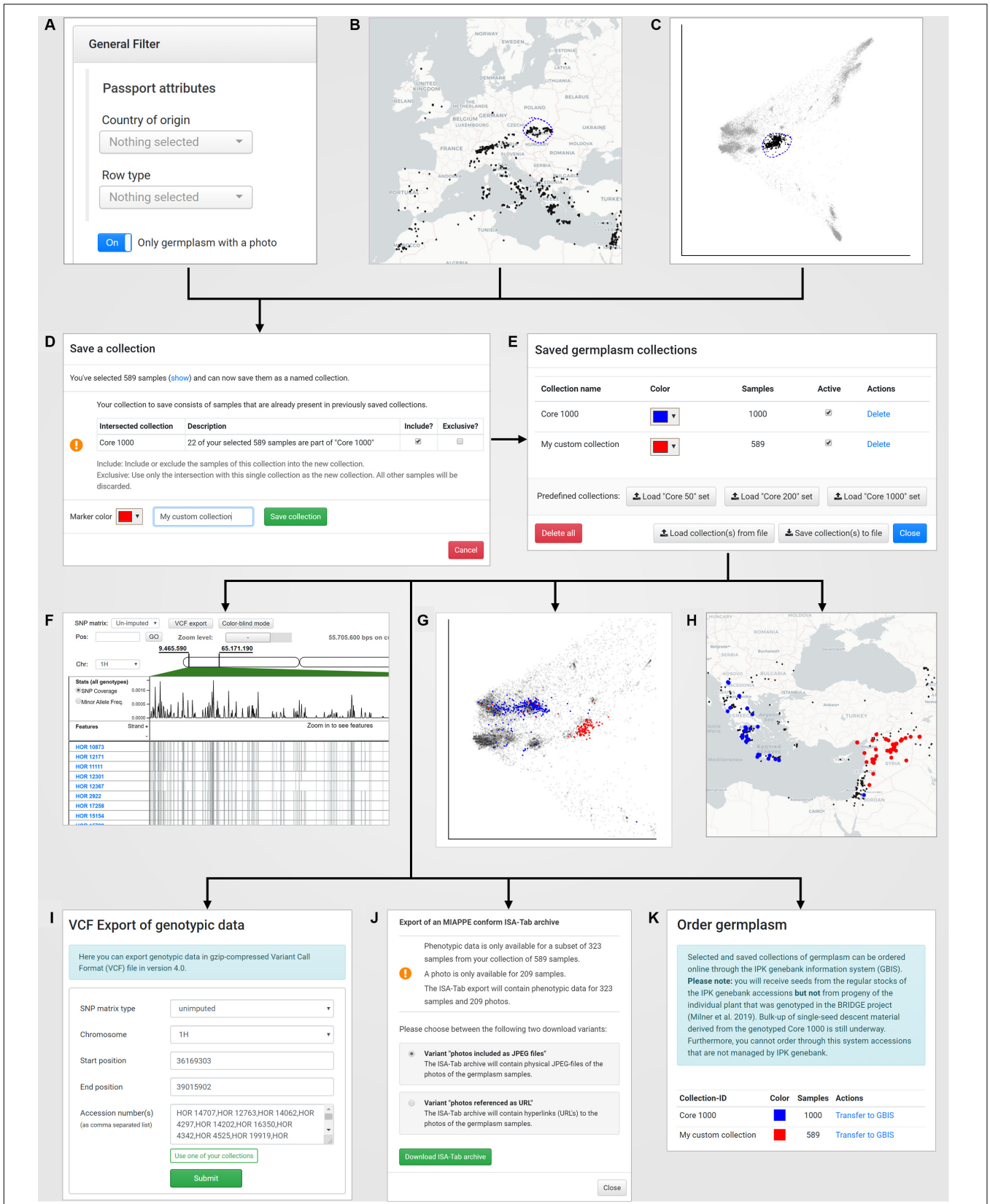

FIGURE 4 | Continued

FIGURE 4 | The concept of "named collections" in combination with the visual analytics concept of "interactive brushing and linking:" A new germplasm collection can be created by **(A)** defining filters on passport records (under "Search Germplasm"), **(B)** a lasso selection in the world map (under "Geographic Origins"), **(C)** a lasso selection in a PCA plot (under "Genetic Diversity"). The save dialog **(D)** allows to store the current collection with a custom name und tagging color. Furthermore, the save dialog automatically detects intersections of the current selected germplasm samples with already existing germplasm collections and provides a function to add, subtract or intersect the sample lists. After saving, the new collection is available in the "Saved germplasm collections" dialog that is located in the menu "Collections" **(E)** and can be reused for application-wide visualization in the SNP browser **(F)**, the PCA plots for exploration of the genetic diversity space **(G)** and the world map **(H)**. Also, the SNP data for that collection can be exported to a VCF file **(I)**, the phenotypic and passport data including pictures of the selected accessions in the collection can be downloaded as a MIAPPE compliant ISA-Tab archive **(J)** and finally the germplasm for the collection can be ordered online from the IPK genebank information system **(K)**.

TABLE 1 | Numbers of germplasm samples with specific available data attributes.

Available data	Number of data records
Germplasm samples with passport data	22,626
Genotyped germplasm samples (available in VCF file)	22,621
Germplasm samples with at least one observed phenotype	9,527
Germplasm samples with spike photographs	6,162
Germplasm samples with geographical (GPS) coordinates	2,862

TABLE 2 | Available passport data attributes with corresponding MCPD codes (Multi-Crop Passport Descriptor Codes) (Alercia et al., 2015) and data completeness.

Passport attribute	MCPD code	Data completeness
Country of origin	ORIGCTY	88.64%
Subtaxon/Subspecies	SUBTAXA	99.99%
Biological status of accession	SAMPSTAT	99.57%
Donor institute	DONORCODE	N/A
Location of collecting site	COLLSITE	29.27%
Full botanical name	N/A	95,15%
Annual growth habit (barley specific)	N/A	90.21%

For the attribute "Donor institute" the specification of data completeness makes no sense and is therefore not provided.

Functionalities and Features

BRIDGE consists of data domain specific search, visualization and export modules. Their functionalities and features will be described in the following.

Sample Collection Manager

The sample collection manager (**Figure 4E**) stores user-defined germplasm collections with a title and a specific color that is later used as a tagging color in the data visualization modules (**Figures 3C**, **4D–F**). There are predefined collections such as the "Core 50," "Core 200," and "Core 1000" collections as described in Milner et al. (2019) can be loaded directly via dedicated buttons. Each collection can be toggled by a checkbox to show or hide the collection in all visualization modules simultaneously. All collections are saved for private access only in the IndexedDB of the user's browser and persist after leaving the web portal and even computer shutdowns (Kimak and Ellman, 2015), unless the private browsing options are used. This approach makes it possible to avoid server-side storage of potentially sensitive data and a user authentication mechanism that would require prior registration. The individual set of collections can also be exported to the local filesystem as a JSON-document and reversely be loaded again.

Passport, Phenotypic, and Variant Data Based Germplasm Search

BRIDGE provides a combined search functionality for passport, phenotypic, and genotypic data. Users are able to filter for any combination of passport, phenotypic or variant data (SNPs) or a combination of all three data domains (**Figure 3A**). To avoid complex database queries involving computation-intensive SQL joins in the database backend, the three data domains are queried separately, and an intersection of the sample IDs of all three result sets is then sent to the frontend. The intersection is the result set that fulfills the search conditions on all three data domains. In addition, the direct search for germplasm via the accession number is possible. The search result can be exported as a CSV file.

SNP Browser

The SNP browser, first described in Basterrechea (2017), allows the visual exploration of the SNP matrices over all sequenced germplasm samples (**Figure 3B**). The user interface consists of three parts. The first part with the navigation in the upper area displays the current genome coordinates and allows to change them. The displayed SNP matrix can be changed via a dedicated selector. Furthermore, there are buttons to zoom in and out, to trigger a VCF export and to toggle a "colorblind mode" for people with red-green visual impairment. The second part below the navigation has a variable display area whose display content depends on the current zoom level. At the maximum zoom level, the nucleotides of the reference genome are displayed. When zooming out, the display changes to show SNP density, minor allele frequency (calculated across all germplasm samples) and a track with genes in genomic windows of variable size. The third part in the lower area of the user interface shows the genotypes of the currently selected collection as tracks. At maximum zoom level each track shows the distinct variants for all marker positions in the manner of a slice of the SNP matrix. When zooming out, the track visualization changes to show SNP coverage for each genotype in the current genomic window. The user can move the viewport and thus the sample list and genomic window by dragging with the mouse pointer. The default zoom level is the nucleotide level view where one can see single SNPs in a linear arrangement. The export of SNP matrices in variant call format (VCF) (Danecek et al., 2011) can be directly triggered from the current viewport of the SNP browser. Then the current list of germplasm samples

TABLE 3 | Observed phenotypic traits of barley spikes and number of accessions per trait value in the study panel.

Phenotypic trait	Possible values	# Accessions
Row-type convarity	convar. *vulgare*/convar. *hexastichon*	5,572
	convar. *distichon*	3,030
	convar. *deficiens*	411
	convar. *labile*	293
	convar. *intermedium*	207
Spike density	Lax	7,804
	Middle	1,483
	Dense	234
Spike branching	Unbranched	9,367
	Branched	153
Spike brittleness	No	9,517
	Yes	4
Grain hull / Cover of caryopses	Covered grain	8,630
	Naked grain	882
Color of naked seeds	Yellow	726
	Between black and brown	57
	Black	50
	Green	32
	Purple	8
	Other	9
Length of rachilla hairs	Long	6,019
	Short	3,474
Awns roughness	Rough	8,926
	Smooth	421
Presence of awns central spikelet	Awnless or very short awns (tips)	71
	Awns short (up to spike length)	450
	Awns long (1.5–3 times spike length)	8,850
	Sessile hoods (sessile or on short)	27
	Elevated hoods (Hood over 1 cm long awns shaped stems)	19
	Hoods with end awn	19
	Elevated hoods with end awn	34
Presence of awns lateral spikelets	Awnless or very short awns (tips)	203
	Awns short (up to spike length)	419
	Awns long (1.5–3 times spike length)	5,276
	Sessile hoods (sessile or on short)	19
	Elevated hoods (Hood over 1 cm long awns shaped stems)	17
	Hoods with end awn	15
	Elevated hoods with end awn	20
Presence of awns on glumes	Awnless or very short awns (tips)	8,725
	Awns short (up to spike length)	785
	Awns long (1.5–3 times spike length)	10
	Sessile hoods (sessile or on short)	0
	Elevated hoods (Hood over 1 cm long awns shaped stems)	0
	Hoods with end awn	0
	Elevated hoods with end awn	1
Glume width	All glumes are narrow (<1 mm)	5,959
	All or some glumes are broad (1–2 mm)	56
Color of glumes	Yellow	5,614
	Gray	122
	Black	115
	Brown	92
	Purple	71
	Green	1
	Other	4

and the genome coordinates will be used as parameters for the VCF export. The variant data is stored in an ORACLE RDBMS (version 12c) using columnar partitioning storage and compressed in-memory-population to provide fast access to the marker data from the SNP browser frontend. The backend was implemented with Grails (Smith and Ledbrook, 2009) in version 3.3.8 and the frontend was implemented with ReactJS[1] in version 16.2.0. The communication between both was realized with a RESTful API.

Interactive Scatterplots for Results of Dimensionality Reduction

Exploring population structure and its correlates by dimension reduction methods such as PCA is a commonly used method in population genetics. BRIDGE includes an interactive visualization of the genetic diversity space defined by PCA and t-SNE, allowing selection of dimensions for display, zooming, lasso selection and highlighting accessions according to their country of origin or phenotypic attributes (**Figure 3C**). The precalculated results of dimensionality reduction methods are stored in the Oracle RDBMS. A lasso selection of germplasm samples inside the scatterplots automatically highlights the countries of origin on the world map and aggregates counts and percentage shares for each country in a tabular summary. Conversely, it is possible to highlight the respective germplasm samples in the diversity space based on a selection of a single country or multiple countries. This enables to inspect the population structure and geographic distribution in an explorative and interactive manner. Color schemes of the scatterplots can be changed by the user according to three options. The first color scheme option uses the user-defined tagging colors of the particular saved germplasm collections. Germplasm, which is not included in any collection, is colorized black. The second color scheme option colorizes the germplasm samples according to their Hamming distance (Hamming, 1950) to the Morex reference genome based on the SNP data. The third option colorizes the dots according to the values of phenotypic traits. The user interface provides a control handle to change the currently colorized trait. The color scheme can be changed by the user to meet personal preferences. This colorization mode handles ordinal and interval scaled trait variables. In addition,

[1] https://reactjs.org

the values of the phenotypic traits are displayed on the world map summarized by individual countries. The scatterplot matrix visualization allows the inspection of the first four principal components of PCA results while providing the ability for lasso selections and the highlighting of samples according to their collection tagging color (**Figure 3F**).

Interactive World Map

The interactive political world map provides a geographical visualization for those genotypes that have geographical coordinates of their collection site available in their passport data (**Figure 3D**). Furthermore, it allows a lasso selection of genotypes, which can be saved as a named germplasm collection. Vice versa, germplasm samples included in a saved germplasm collection are automatically highlighted with their respective tagging color.

Interactive Manhattan Plots of Genome Wide Association Studies (GWAS)

The correlation between phenotypic traits and genotypic information can be explored in the zoomable Manhattan plot visualization (**Figure 3E**). By clicking on a data point of a variant *p*-value the application automatically jumps to that variant position in the embedded SNP browser. The trait to be displayed can be switched by a select box.

VCF Export

This module allows the parameterized export of a subset of the underlying SNP matrices (**Figure 4I**). The export parameter form consists of input fields for the SNP data type, chromosome, start position, end position and accession numbers. It can be completed manually or filled with the corresponding parameters of the current viewport of the SNP browser. For one or more genotypes, markers within a single chromosome can be exported as a VCF file in version 4.0 (Danecek et al., 2011). The VCF export is limited to 1,000,000 data points due to performance reasons. This allows for example to export 1,000 SNP's for 1,000 genotypes or 100,000 SNP's for 10 genotypes.

MIAPPE-Compliant ISA-Tab Export

Beside the export of genetic information via VCF files, it is also possible to export the phenotypic data records in the ISA-Tab format (Sansone et al., 2012), which combines a machine-readable and a textual human-readable representation (**Figure 4J**). The observation scores used for phenotyping are described in an embedded "Trait Definition File." Thus, the provided metadata complies with the version 1.1 of the MIAPPE standard (Krajewski et al., 2015; Ćwiek-Kupczyńska et al., 2016; Papoutsoglou et al., 2020). Due to the strong diversity of phenotypic research data, the standard was initially developed to explain the minimal information that is necessary to describe plant phenotypic experiments. MIAPPE is still under active development[2]. For a more convenient use, BRIDGE offers the possibility to choose between a full data export, which includes the corresponding digital spike image files physically and a metadata-only download with persistent URLs to access the images online.

Germplasm Order Service

Saved germplasm collections can directly be transferred to the IPK genebank information system (GBIS) (Oppermann et al., 2015) to get further information or order seeds of accessions of interest (**Figure 4K**). The transfer is implemented via an HTTP call to the GBIS RESTful API, which checks whether a sufficient amount of seeds is stored at IPK genebank and is available for distribution under the terms of the IPK Genebank Material Transfer Agreement (SMTA). Accessions with insufficient numbers of seeds on stock are automatically excluded but displayed as a list to the user including the reasons for exclusion (e.g., limited seed stock). Users can then fill out the order form in GBIS to receive the desired seed material. We would like to note, that single-seed descent material used throughout the BRIDGE experiments cannot be ordered at present. Material will then be procured from the regular stocks of the IPK genebank. Seeds from accessions that are not maintained at IPK genebank [e.g., accessions from the Swiss and Chinese genebank included in Milner et al. (2019) study] cannot be ordered through GBIS but need to be placed to the respective genebank.

Interactive Help

One important task when designing a graphical user interface (GUI) is the conception of menu elements, search fields, buttons and the general layout. Complex applications like BRIDGE that are comprised of many different modules can be overwhelming for first-time users. We therefore minimized the number of interactive elements in the GUI exposed to the user at any single point in time. Only the main navigation is always visible on the menu to the left and does not contain nested submenus (**Figure 1**). Moreover, we integrated an interactive tutorial that guides the process of discovering the potential and power of the application. This discovery is supported by a context-dependent, visual highlighting of control elements and a textual description of their function. It is implemented with the help of the IntroJS library (Arasteh and Mehrabani, 2013).

RESULTS

BRIDGE allows an easy-to-use access to exploratory data analysis of passport and genotypic data of a worldwide panel of barley accessions, phenotypic data, related diversity data and downstream research results (**Figure 4**). Through the use of techniques that users experienced through popular web pages that use world maps or lasso selection like in image editing programs, the barrier to use BRIDGE is considerably lower than in similar systems. Furthermore, much of BRIDGE has been designed with a novice user in mind and we guide them with our interactive help function. Due to the high integration level of the visualization and analysis tools as a single-page application, data sets representing research results can be conveniently and quickly reviewed and analyzed. In comparison to a traditional multi-page

[2]https://www.miappe.org

application there are no time-consuming page reloads from the web server initiated by actions via navigation links or functional buttons. This leads to an increased user experience when using the application. Hence, the single-page application behaves and feels more like a traditional standalone desktop application with minimal time delay between a user action and the corresponding application response. Another advantage is the prevention of time-consuming and error-prone manual data conversions and data transfers between multiple standalone programs for each data visualization domain.

The identification of plant genotypes that meet certain criteria for adaptation to climatic and agronomic conditions as well as criteria for nutritional traits is a basic requirement for successful cultivation and improvements in breeding. This challenge is multidimensional, as multiple criteria must be met by a single crop genotype. Digital information systems such as BRIDGE can support this identification of suitable crop genotypes by providing convenient access and searchability in this multidimensional data space. Plant scientists might be interested in finding candidates of barley accession with specific genetic variants for a research topic they are working on, while barley breeders are looking for high diversity of target quality traits to complement their breeding panel with promising plant genetic resources. These different user groups have distinct demands for their workflows and subsequently we show how BRIDGE can tailor these demands by presenting selected exemplary use cases. Nevertheless, this is by far not a complete set what can be done theoretically by using the application.

Exemplary Use Cases

Finding Accessions With Characteristic Genotypic and Phenotypic Features

BRIDGE allows the search of germplasm with specific genotypic and/or phenotypic traits. By using the variant filter feature in the "Search germplasm" panel, the user can find germplasm with specific variants for one or more SNPs. The filter setting on variants can be combined with filters for passport data and phenotypic attributes of interest (**Figure 1**, available under the feature "Search Germplasm" and one starting point to create a user specific sample collection, see **Figure 4E**). This feature might be useful for breeders as well as for scientists that want to identify genotypes matching certain criteria for crop improvement or similar research questions. Another functionality is the ability to find the corresponding genomic region for a given gene of interest as a function of the integrated SNP browser. A scientist or breeder can then export marker data as VCF files for his germplasm samples of interest to feed this variation data into subsequent analysis steps (**Figure 4I**).

Ordering Germplasm With Specific Attributes of Interest

Through the integration with the IPK Genebank Information System (GBIS), it is possible to order germplasm of interest by transferring individually created germplasm collections to GBIS (**Figures 4E,D,K**). This is convenient if a user wishes to order, for example, germplasm from one of the predefined core sets or a user-defined collection of germplasm samples identified by prior exploratory data analysis. The integration avoids the manual transfer of germplasm sample lists through export and import, thus reducing the probability for errors.

Usage as a Decision Support System for Genebank Data Curation

Mascher et al. (2019) gave one example how BRIDGE was used as a decision support system to improve the data quality of the barley collection of the IPK genebank. The combined and interactive access to passport data and results of genetic clustering algorithms revealed that thousands of Ethiopian accessions in the IPK genebank had a false biological status of "wild" instead of the correct one: "domesticated." This can be verified in BRIDGE by searching for germplasm with "Ethiopia" as country of origin and with "wild" as biological status (defining filters using the "Search Germplasm" feature, see **Figures 1**, **4A**). The search result must be saved as a user-defined collection (**Figures 4D,E**). When switching to the PCA plot, it is clearly visible that these wild Ethiopian accessions are located in a cluster of domesticated barley (for filter definition using the feature "Search Germplasm" and for visualization the features "PCA Matrixplot" and "Geographic Origins," see in **Figure 1** and similar visualized like in **Figures 4G,H**).

Data Stewardship as Necessary Precondition for Exploratory Data Analysis

An integrated exploratory analysis of phenotypic and genotypic data is possible if used plant material and probes have been consequently processed in a FAIR-compliant laboratory process. In BRIDGE, we applied homogenized protocols for phenotyping, genotyping, data analysis and storage tracking, accompanied by a strict data stewardship. This procedure enables the return of investment of the efforts for data quality management, which is essential for the integrated and explorative analysis of primarily heterogeneous data. In addition, the iterative process of exploratory data analysis enables the continuous support of data curation within the framework of data stewardship principles (Wilkinson et al., 2016; **Figure 5**).

Comparison With Similar Web Portals

Integrative applications for web-based exploration of data in the data domain of genotyping and phenotyping data are scarce especially those focussing on plant genetic resources. In the following, we compare BRIDGE to two well-established barley specific web portals (Germinate3 and T3/Barley) and present the individual strengths and weaknesses. Results are summarized in **Table 4**. We note that there are more systems available, especially when considering other organisms, and this list should not be regarded as complete.

Germinate3

One application framework for which a barley instance is available is Germinate3 (Shaw et al., 2017). Compared to BRIDGE, it shares some of the features and behaves more like a traditional table-based data warehouse than a visual analytics driven application. Thus, the individual tables do not interact with each other and can be regarded as static instances. It

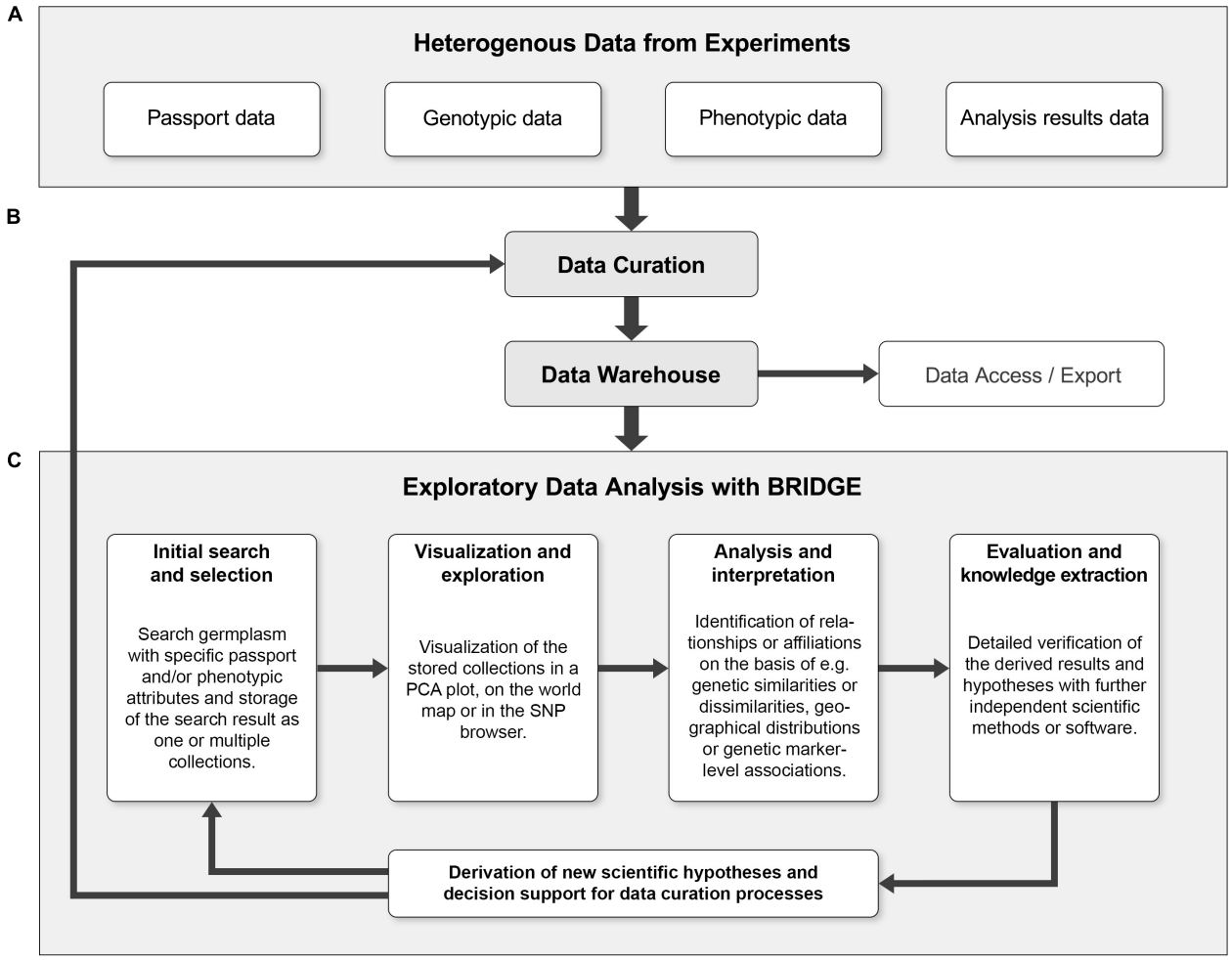

FIGURE 5 | The combination of exploratory data analysis and continuous data curation can form a cycle that leads to continuous quality improvements. **(A)** The heterogeneous primary data from genotyping and phenotyping experiments combined with data of further analysis steps like PCA or GWAS ("Analysis results data") are collected and serve as initial data sources. **(B)** The input data is subjected to a first curation step and then fed into the data warehouse. The data warehouse allows programmatic access to the data and export of the data in standardized formats. **(C)** The iterative process of exploratory data analysis (EDA) makes it possible to derive new scientific hypotheses without prior assumptions. Furthermore, the result of an EDA iteration can reveal data inconsistencies and thus be a starting point for a subsequent data curation step. Ideally, these processes form a continuous cycle, which can lead to continuous quality improvement of primary research data and derived data. Using an entry point such as the germplasm search, a "named collection" of germplasm to be visualized or explored is created. Users can then search for outliers, clusters or other visual keys in a genetic diversity plot (such as PCA). When a sample or a set of samples of specific interest is identified and highlighted by a lasso selection, it is possible to save this subset as a new collection and look for further details, e.g., via the SNP profiles in the SNP browser. This detailed look into the data could be enough to find something unusual that might be worth further experimentation or lead to new scientific findings.

also follows the principle of an integrated and uniform user interface and the concept of sample collections, called "groups." A functionality shared with BRIDGE is the geographical world map with the possibility to pick germplasm samples into a collection by drawing a line around them. A combined search for passport and phenotypic data as well as a seamlessly integrated possibility to visualize marker data is not implemented. However, marker data can be exported to different file formats with one being the proprietary Flapjack file format (Milne et al., 2010). This workflow decelerates the speed for cycles of exploratory data analysis because the data has to be downloaded, then imported into the standalone Flapjack software and finally has to be analyzed until new knowledge can be generated. This in turn decreases the user experience and is prone to produce errors from either shifting platforms or from version conflicts. Germinate3 also provides the ability to render scatterplots which is available for phenotypic data. Furthermore, there is no possibility to export genotypic and phenotypic data in standardized file formats. A possibility to visualize the results of GWAS by Manhattan diagram is not implemented. Nevertheless, Germinate3 can be described as a generic database solution for plant genetic resources. Several different plant species have been set up including barley, chickpea, eggplant, maize and wheat (with others in active development). While BRIDGE offers interactive visual analytics tools Germinate3 is more suited for users that prefer table-based analysis.

TABLE 4 | Comparison between BRIDGE, Germinate3, and T3/Barley.

Resource name	Resource type	Custom collections	Interactive brushing and linking	Integrated visualizations	Interactive plots	Data export
Germinate3	Data warehouse	(✓)	×	✓	(✓)	CSV, Flapjack
T3/Barley	Data warehouse	×	×	✓	×	CSV, Flapjack
BRIDGE	Visual analytics webtool	✓	✓	✓	✓	ISA-tab, VCF, CSV

The features of the resources are compared based on: resource type, the ability to subset the data into custom collections, if the concept of "interactive brushing and linking" has been implemented, the ability to generate graphical results, the interactivity of generated plots, and the data export formats. The "✓" shows a full compliance, while "(✓)" shows a partial compliance and the "×" shows no compliance.

T3/Barley

T3/Barley[3] is the successor of The Hordeum Toolbox (Blake et al., 2012). It has a multitude of functions and a complex user interface with a main navigation consisting of several submenus. Even though the portal offers a lot of functionality, the user is overwhelmed by many acronyms and abbreviations that are not explained. For the visualization of genetic diversity data, the portal offers an export function to create files for the standalone genome viewer Flapjack (Milne et al., 2010). Data export is possible via CSV formats which makes downstream analysis challenging as it is the user's responsibility to transform the data into appropriate standard formats. In contrast to BRIDGE, users of the T3/Barley resource need a deep understanding of the system. New users of this resource may have difficulty obtaining meaningful results because many tools and visualizations are hidden behind layers of parameter selection screens.

DISCUSSION

In the following, we will discuss the key benefits of information systems and tools such as BRIDGE in the context of the growing volume of heterogeneous genomic and phenomic data, the integration of new data sets and possible useful functional enhancements.

Integration and Information Retrieval of Growing Genomic and Phenomic Data Sets

Technology breakthroughs of the last decade like next generation sequencing and high-throughput phenotyping have significantly changed the scientific landscape (Elshire et al., 2011; Pieruschka and Schurr, 2019). The speed and amount of data generation was accelerated to a new level and it is becoming more and more challenging to analyze and visualize scientific findings in a comprehensive way. Many software tools have been developed in recent years with similar goals: Frustration-free access to constantly growing life science data sets for exploration, analysis and interoperability. As systems for the integration and linkage of heterogeneous data sets, they can greatly help to avoid the problem of not seeing the forest for the trees and offer a low-barrier assistance to extract hidden treasures from the "sea of bioinformatics data" (Roos, 2001), a challenge that has existed for at least 20 years. Furthermore, such web portals can serve as a unified entry point for validation, recapitulation and export of PGR data sets as well as an "exploratory data analysis playground" for generating new research ideas and hypotheses (de Mast and Kemper, 2009).

In addition, application concepts such as BRIDGE can be regarded as proof-of-concepts and blueprints for the software-supported transformation of genebanks into bio-digital resource centers, thus helping to close the gap between the preservation of crop diversity in genebanks and plant breeding (Mascher et al., 2019). This support can be achieved by leveraging the power of visual analytics and exploratory data analysis to keep track of the ever-growing volume of PGR related data sets. We highlighted one use case, where such exploratory data analysis could be applied to directly increase the data quality by correcting falsely labeled wild germplasm. Thus, processes need to be defined which can directly annotate or change the passport information in information systems managing PGR data. In the future, it will be possible in BRIDGE to switch between either historical passport data or passport data updated by the continuously ongoing data curation process.

Integration of New Data Sets

As more and more large collections of genetic material are being molecularly characterized using NGS technology (Weigel and Mott, 2009; The 3,000 rice genomes project, 2014; Mascher et al., 2019), one possible step is to integrate genomic data of germplasm from other genebanks. This would increase the overall allelic diversity of the entire barley panel and thus increase the usefulness in finding genotypes with specific or rare variants. Many challenges, have yet to be addressed to fully integrate different data sets. Currently, variants are called in regards to a single reference genome and will only record polymorphisms detected in the analyzed population. An integration of several such data sets could lead to a contrasting result for individual variants, with a certain variant only recorded in one population while another variant is reported in another. In addition, the version of reference genome sequence would have to be identical in order to compare or integrate different data sets.

Since a single reference genome sequence cannot explain all possible variants in a given species, researchers have started to study variants on a larger scale with regard to the so-called pan-genome. It was previously reported that much variation in elite breeding material has been lost due to the domestication bottleneck (Tanksley and McCouch, 1997; Hyten et al., 2006; Kilian et al., 2006; Zhu et al., 2007). Especially in resistance breeding, the use of PGR or plant

[3]https://triticeaetoolbox.org/barley

wild relatives in breeding programs is a promising field for the introduction of new resistance genes. Additional reference-quality genome sequences will allow researchers and plant breeders to better exploit molecular marker data in diverse germplasm. We imagine the possibility to easily switch the reference genome sequence in the SNP browser of BRIDGE or even the integration of a pan-genome browser will help with such an analysis.

Another aspect could be the inclusion of sequences from other sequencing techniques such as exome capture resequencing (Mascher et al., 2013) or WGS data from genotypes already used in BRIDGE to increase sequencing depth. Such an increase would be beneficial for multiple objectives, such as reducing the proportion of missing data and the ability to call more variants. In the case of WGS data this would also allow comparative analysis of structural variations (Zhou et al., 2018). At the time of writing, it is planned to sequence between 100 and 1000 genotypes belonging to one of the core sets for BRIDGE (Monat et al., 2019). These accessions will be subjected to WGS and will help to further characterize the genomic diversity of barley germplasm.

Useful Feature Enhancements

We plan to improve the current implementation of the SNP browser. One feature we believe requires such an improvement is the feature track. Currently, only the gene boundaries are drawn and a mouse overlay displays the gene name. It would be useful to see the intron-exon structure as well as different isoforms of genes to determine if a variant could have an effect on the coding sequence. In addition, changes to the protein sequence should be visible at a glance. The feature track could also be used to highlight QTL regions where users can examine the annotated genes and search for causal candidates for an observed phenotype. Such a QTL region would be defined by a collection of SNP markers. A lasso selection in a GWAS Manhattan plot or a range selection in the SNP browser could be a starting point to generate such a collection. We are also considering improving the functionality of the statistics track. At the moment, both the SNP coverage and the minor allele frequency are precalculated for the entire data set. However, it would be useful to dynamically calculate these metrics for a selected germplasm subset. Furthermore, the differences in metrics between the entire data set and a collection might be helpful, and we are experimenting with different methods to visualize that. It would also be very useful to be able to group similar SNP profiles of different genotypes together. This could provide a haplotype based view on the selected collection and a way to reduce visual noise. A very useful feature for the SNP browser would be the ability to dynamically filter variants according to various criteria like e.g., minor allele frequency, number of heterozygous calls, quality of the SNP and number of supporting reads.

We also want to simplify the handling of germplasm collections throughout BRIDGE. Currently, only one list of germplasm collections can be managed at a time. A potential worksheet feature would be helpful to be able to manage several different lists at the same time. This feature would allow the organization of multiple and different sets of sample collections to address different research topics or explorative workflows. Concerning the germplasm search, we hope to extend this function to include genotyping data that allows a combined search not only for passport and phenotypic data, but also for genotypic data.

Another interesting and promising feature would be an online calculation for dimensionality reductions, such as PCA or t-SNE on the SNP data. Based on an individually generated subset of germplasm samples, this approach would allow for a semi-automated exploration of the population structure at a fine-grained resolution, e.g., by applying a PCA to a local cluster of genotypes. Due to the large size of the SNP data and the resulting computational load, these calculations cannot be performed client-side on the user's device. Instead these calculations would have to be performed on the web server, which in turn would have to be sufficiently powerful in terms of CPU and RAM. Alternatively, a calculation in an elastic cloud environment could be considered that can react flexibly to the required calculation effort. Due to the longer calculation times in general and the potentially parallel occurrence of a calculation request through simultaneous actions by multiple users, a job queue would be required. This queue could process the individual calculations in turn and send the user a notification that the calculation is complete.

CONCLUSION

BRIDGE is an application concept and implementation for the visual analytics driven exploration of data of plant genetic resources (PGR), mainly stored at the German Federal *ex situ* genebank at IPK Gatersleben. We presented the benefits of a quality curated data warehouse of integrated genomics and phenomics data based on a deeply genotyped and phenotyped worldwide barley germplasm collection of 22,626 genotypes. The genotypic and phenotypic data is extended with linked downstream analysis data like GWAS results and dimensionality reduction results of the SNP data. In particular, we demonstrated the benefit of multiple entry points for germplasm search, analysis, knowledge extraction and data export allowing plant scientists and plant breeders to extract domain specific information of personal interest. Users can benefit from this curated combination of legacy and newly derived data by the ability to order barley accessions of specific relevance and by incorporating these ordered PGRs in their own research or breeding efforts. Furthermore, the user can benefit by exporting subsets of the provided data in common file formats like VCF for data of genetic variants or MIAPPE-compliant ISA-Tab archives for phenotypic data. Application concepts like BRIDGE can act as a proof-of-concept for the software-aided transformation of genebanks into bio-digital resource centers allowing to close the gap between the conserved crop diversity, plant breeding and research. Moreover, they can serve as a first entry point for data curation and scientific hypothesis generation. Feedback from IPK and close collaboration partners has revealed the potential

for a number of meaningful and useful possibilities and ideas for extending the functionality of the system, which will be implemented in future versions of BRIDGE.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.nature.com/articles/s41588-018-0266-x#data-availability, https://doi.org/10.5447/IPK/2018/9, and https://doi.org/10.5447/IPK/2018/10.

AUTHOR CONTRIBUTIONS

NS, MM, and US designed the study. NS supervised the experiments. MM analyzed the data. PK designed and developed the main parts of the application software. MB, MM, and ML designed and MB implemented the SNP browser. DS, PK, SB, ML, and DA curated, analyzed, and imported data. ML developed the data storage concept. DA implemented the ISA-Tab export API. US and ML supervised the development of the portal. PK, SB, DA, US, and ML wrote the manuscript with contributions from all co-authors. All authors read and approved the final manuscript.

FUNDING

We gratefully acknowledge funding from the Leibniz Association to US, MM, and NS (Pakt für Forschung und Innovation: SAW-2015-IPK-1 BRIDGE); from the German Ministry of Education and Research (BMBF) Grant No. 031A536A; de.NBI to US and Grant No. 031B0190A; SHAPE to NS, US, and MM.

ACKNOWLEDGMENTS

We thank J. Bauernfeind, T. Münch, and H. Miehe for administration of the IT infrastructure; M. Oppermann, S. Weise, M. Ullrich, and H. Knüpffer for providing the interface to the IPK Genebank Information System; A. Graner and J. C. Reif for stimulating discussions.

REFERENCES

Alercia, A., Diulgheroff, S., and Mackay, M. (2015). *FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1]*. Available online at: https://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/ (accessed January 20, 2020).

Arasteh, E., and Mehrabani, A. (2013). *Instant Introjs Learn How To Work With The Introjs Library To Create Useful, Step-By-Step Help And Introductions For Websites And Applications*. Birmingham: Packt Pub.

Basterrechea, M. (2017). *Web-Interface to Browse, Filter And Visualize Plant Genotyping Data*. Available online at: http://lup.lub.lu.se/student-papers/record/8910613 (accessed January 25, 2020).

Blake, V. C., Kling, J. G., Hayes, P. M., Jannink, J.-L., Jillella, S. R., Lee, J., et al. (2012). The hordeum toolbox: the barley coordinated agricultural project genotype and phenotype resource. *Plant Genome J.* 5:81. doi: 10.3835/plantgenome2012.03.0002

Bolger, A. M., Poorter, H., Dumschott, K., Bolger, M. E., Arend, D., Osorio, S., et al. (2019). Computational aspects underlying genome to phenome analysis in plants. *Plant J.* 97, 182–198. doi: 10.1111/tpj.14179

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and Sequence Database Collaboration (2016). The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 44, D48–D50. doi: 10.1093/nar/gkv1323

Coppens, F., Wuyts, N., Inzé, D., and Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr. Opin. Syst. Biol.* 4, 58–63.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., et al. (2016). Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12:44. doi: 10.1186/s13007-016-0144-4

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.

de Mast, J., and Kemper, B. P. H. (2009). Principles of exploratory data analysis in problem solving: What can we learn from a well-known case? *Qual. Eng.* 21, 366–375. doi: 10.1080/08982110903188276

De Volder, K. (2006). *JQuery: A Generic Code Browser With A Declarative Configuration Language*. Berlin: Springer.

Dirks, R., van Dun, K., de Snoo, C. B., van den Berg, M., Lelivelt, C. L. C., Voermans, W., et al. (2009). Reverse breeding: a novel breeding approach based on engineered meiosis. *Plant Biotechnol. J.* 7, 837–845. doi: 10.1111/j.1467-7652.2009.00450.x

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

FAO (2018). *Building Climate Resilience For Food Security And Nutrition*. Rome: FAO.

Filipova, O. (2016). *Learning Vue. js 2*. Birmingham: Packt Publishing Ltd.

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291.

González, M. Y., Philipp, N., Schulthess, A. W., Weise, S., Zhao, Y., Börner, A., et al. (2018). Unlocking historical phenotypic data from an ex situ collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.). *Theor. Appl. Genet.* 131, 2009–2019. doi: 10.1007/s00122-018-3129-z

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Syst. Tech. J.* 29, 147–160. doi: 10.1002/j.1538-7305.1950.tb00463.x

Hazzard, E. (2011). *Openlayers 2.10 Beginner's Guide*. Birmingham: Packt Publishing Ltd.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26, 121–135. doi: 10.1105/tpc.113.119982

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat. Rev. Genet.* 11:855.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., et al. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U.S.A.* 103, 16666–16671. doi: 10.1073/pnas.0604379103

IBSC (2016a). *Functional Information Of High Confidence Genes Of Barley cv. Morex. 5.1 MB*. Lawndale, CA: IBSC, doi: 10.5447/IPK/2016/45

IBSC (2016b). *Structural Information Of Low-Confidence Genes Of Barley cv. Morex. 114.9 MB*. Lawndale, CA: IBSC, doi: 10.5447/IPK/2016/46

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740. doi: 10.1186/1471-2164-15-740

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph.* 8, 1–8. doi: 10.1109/2945.981847

Kilian, B., Özkan, H., Kohl, J., von Haeseler, A., Barale, F., Deusch, O., et al. (2006). Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol. Genet. Genomics* 276, 230–241. doi: 10.1007/s00438-006-0136-6

Kimak, S., and Ellman, J. (2015). "The role of HTML5 IndexedDB, the past, present and future," in *Proceedings of the 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, London, 379–383.

Krajewski, P., Chen, D., Æwiek, H., van Dijk, A. D., Fiorani, F., Kersey, P., et al. (2015). Towards recommendations for metadata and data handling in plant phenotyping. *J. Exp. Bot.* 66, 5417–5427. doi: 10.1093/jxb/ery006

Leff, A., and Rayfield, J. T. (2001). "Web-application development using the Model/View/controller design pattern," in *Proceedings Fifth IEEE International Enterprise Distributed Object Computing Conference*, Seattle, WA.

Maaten, L., and van der Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A., Clissold, L., Sampath, D., et al. (2013). Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* 76, 494–505. doi: 10.1111/tpj.12294

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* 51, 1076–1081. doi: 10.1038/s41588-019-0443-6

Miles, A., Ralph, P., Rae, S., and Pisupati, R. (2019). *Cggh/Scikit-Allel: v1.2.0*. London: Zenodo.

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., et al. (2010). Flapjack–graphical genotype visualization. *Bioinformatics* 26, 3133–3134. doi: 10.1093/bioinformatics/btq580

Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., et al. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51, 319–326. doi: 10.1038/s41588-018-0266-x

Monat, C., Schreiber, M., Stein, N., and Mascher, M. (2019). Prospects of pan-genomics in barley. *Theor. Appl. Genet.* 132, 785–796. doi: 10.1007/s00122-018-3234-z

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 17:53. doi: 10.1186/s13059-016-0961-9

Nagel, M., Vogel, H., Landjeva, S., Buck-Sorlin, G., Lohwasser, U., Scholz, U., et al. (2009). Seed conservation in ex situ genebanks—genetic studies on longevity in barley. *Euphytica* 170, 5–14.

NCBI Resource Coordinators (2017). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 45, D12–D17. doi: 10.1093/nar/gkw1071

Oppermann, M., Weise, S., Dittmann, C., and Knüpffer, H. (2015). GBIS: the information system of the German Genebank. *Database* 2015:bav021. doi: 10.1093/database/bav021

Pachauri, R. K., Mayer, L., and Intergovernmental Panel on Climate Change (eds) (2015). *IPCC: Climate Change 2014: Synthesis Report*. Geneva: Intergovernmental Panel on Climate Change.

Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytol.* 16544. doi: 10.1111/nph.16544

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pieruschka, R., and Schurr, U. (2019). Plant phenotyping: past, present, and future. *Plant Phenom.* 2019:7507131.

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., et al. (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26, 2354–2356. doi: 10.1093/bioinformatics/btq415

Roos, D. S. (2001). Bioinformatics–trying to swim in a sea of data. *Science* 291, 1260–1261. doi: 10.1126/science.291.5507.1260

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126. doi: 10.1038/ng.1054

Shaw, P. D., Raubach, S., Hearne, S. J., Dreher, K., Bryan, G., McKenzie, G., et al. (2017). Germinate 3: development of a common platform to support the distribution of experimental data on crop wild relatives. *Crop Sci.* 57:1259. doi: 10.2135/cropsci2016.09.0814

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. doi: 10.1038/s41586-019-1120-8

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26:1135.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., et al. (2017). *Plotly: Create Interactive Web Graphics Via 'Plotly. js.'. R Package Version 4*.

Silvester, N., Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., Cleland, I., et al. (2018). The european nucleotide archive in 2017. *Nucleic Acids Res.* 46, D36–D40. doi: 10.1093/nar/gkx1125

Smith, G., and Ledbrook, P. (2009). *Grails in Action*. Shelter Island, NY: Manning Publications Co.

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Tateno, Y. (2002). DNA data bank of japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* 30, 27–30. doi: 10.1093/nar/30.1.27

The 3,000 rice genomes project (2014). The 3,000 rice genomes project. *GigaScience* 3:7. doi: 10.1186/2047-217X-3-7

Ullrich, S. E. (2010). *Barley: Production, Improvement, And Uses*. Hoboken, NJ: John Wiley & Sons.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10, 621–630.

Varshney, R. K., Nayak, S. N., May, G. D., and Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530. doi: 10.1016/j.tibtech.2009.05.006

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154.

Voytas, D. F., and Gao, C. (2014). Precision genome engineering and agriculture: opportunities and regulatory challenges. *PLoS Biol.* 12:e1001877. doi: 10.1371/journal.pbio.1001877

Weigel, D., and Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.* 10:107. doi: 10.1186/gb-2009-10-5-107

Weise, S., Oppermann, M., Maggioni, L., van Hintum, T., and Knüpffer, H. (2017). EURISCO: the european search catalogue for plant genetic resources. *Nucleic Acids Res.* 45, D1003–D1008. doi: 10.1093/nar/gkw755

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Zhou, B., Ho, S. S., Zhang, X., Pattni, R., Haraksingh, R. R., and Urban, A. E. (2018). Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J. Med. Genet.* 55, 735–743. doi: 10.1136/jmedgenet-2018-105272

Zhu, Q., Zheng, X., Luo, J., Gaut, B. S., and Ge, S. (2007). Multilocus analysis of nucleotide variation of oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* 24, 875–888. doi: 10.1093/molbev/msm005

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 Plant Science

METHODS
published: 23 June 2020
doi: 10.3389/fpls.2020.00666

Automated Spike Detection in Diverse European Wheat Plants Using Textural Features and the Frangi Filter in 2D Greenhouse Images

Narendra Narisetti[1], *Kerstin Neumann*[2], *Marion S. Röder*[2] *and Evgeny Gladilin*[1]*

[1] Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, [2] Department of Genebank, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

OPEN ACCESS

Edited by:
Jose Antonio Jimenez-Berni, Spanish National Research Council, Spain

Reviewed by:
Suchismita Mondal, International Maize and Wheat Improvement Center, Mexico
Stanley Joseph Miklavcic, University of South Australia, Australia

***Correspondence:**
Evgeny Gladilin
gladilin@ipk-gatersleben.de

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *25 October 2019*
Accepted: *29 April 2020*
Published: *23 June 2020*

Citation:
Narisetti N, Neumann K, Röder MS and Gladilin E (2020) Automated Spike Detection in Diverse European Wheat Plants Using Textural Features and the Frangi Filter in 2D Greenhouse Images. Front. Plant Sci. 11:666. doi: 10.3389/fpls.2020.00666

Spike is one of the crop yield organs in wheat plants. Determination of the phenological stages, including heading time point (HTP), and area of spike from non-invasive phenotyping images provides the necessary information for the inference of growth-related traits. The algorithm previously developed by Qiongyan et al. for spike detection in 2-D images turns out to be less accurate when applied to the European cultivars that produce many more leaves. Therefore, we here present an improved and extended method where (i) wavelet amplitude is used as an input to the Laws texture energy-based neural network instead of original grayscale images and (ii) non-spike structures (e.g., leaves) are subsequently suppressed by combining the result of the neural network prediction with a Frangi-filtered image. Using this two-step approach, a 98.6% overall accuracy of neural network segmentation based on direct comparison with ground-truth data could be achieved. Moreover, the comparative error rate in spike HTP detection and growth correlation among the ground truth, the algorithm developed by Qiongyan et al., and the proposed algorithm are discussed in this paper. The proposed algorithm was also capable of significantly reducing the error rate of the HTP detection by 75% and improving the accuracy of spike area estimation by 50% in comparison with the Qionagyan et al. method. With these algorithmic improvements, HTP detection on a diverse set of 369 plants was performed in a high-throughput manner. This analysis demonstrated that the HTP of 104 plants (comprises of 57 genotypes) with lower biomass and tillering range (e.g., earlier-heading types) were correctly determined. However, fine-tuning or extension of the developed method is required for high biomass plants where spike emerges within green bushes. In conclusion, our proposed method allows significantly more reliable results for HTP detection and spike growth analysis to be achieved in application to European cultivars with earlier-heading types.

Keywords: plant phenotyping, high-throughput analysis, cultivars, spike detection, heading time point (HTP), texture, image segmentation, spike area

1. INTRODUCTION

Wheat is one of the major crop species in the world, with 762 million tons of grain produced annually (FAOSTAT 2018) and providing ≥ 20% of the world's calorie and protein demand (Braun et al., 2010). However, the increasing world population and climate change are major threats to sustainable crop production (Tester and Langridge, 2010). Therefore, concentrated efforts are required to increase crop yield and production to meet future needs. Information-based plant breeding and precision agriculture are fundamental for identifying suitable wheat varieties to increase wheat productivity and production. One of the important components in both crop breeding and precision agriculture is the monitoring of plant developmental growth stages to apply informed-decision-based treatments in field or greenhouse experiments. Phenology influences grain yield components both directly and indirectly (Snape et al., 2001; Zhang et al., 2009), and in this aspect, quantitative assessment of crop phenology plays an important role in precision phenotyping as a quantifier of crop performance.

According to the Feekes scale, wheat growth can be classified into four major growth stages: tillering, stem elongation, heading, and ripening. A more detailed sub-classification is made in the BBCH scale (Witzenberger and Hack, 1989), with BBCH classes 49–59 representing phenology from heading to flowering. The determination of phenological stages is necessary for the interpretation of growth-related traits and stress tolerance acquired from non-invasive phenotyping. It is well-known that the major flowering time gene *PPD-H1* has a direct influence on leaf growth in barley (Digel et al., 2016), and flowering time genes have an impact on abiotic stress tolerance (Habte et al., 2014; Abdel-Ghani et al., 2019). In a study employing non-invasive phenotyping of barley growth, correlation of biomass and tipping time (BBCH49) was high (Neumann et al., 2017) and resulted in a constitutive biomass QTL in the region of *PPD-H1* (Dhanagond et al., 2019). However, tipping time had to be assessed by a time-consuming visual inspection of individual plant images across time. The relationship of biomass to flowering time also holds true for wheat: both crops have delayed flowering in an environment with long growing seasons to allow longer and higher vegetative growth (Cockram et al., 2007). Similar to barley, sensitive or insensitive Ppd-D1 alleles in wheat have been shown to correspond to differences in leaf area (Guo et al., 2018). In winter wheat, an earlier flowering time of semidwarf cultivars was associated with reduced biomass at anthesis (Maeoka et al., 2020). In dryland regions, simulations showed that higher yield derives from an increased biomass before anthesis leading to an increased grain number (Zhao et al., 2019). Non-invasive imaging experiments with a large wheat collection have been conducted to genetically dissect drought and heat-stress tolerance (unpublished data). An automated solution is urgently required for an effective determination of flowering time-related growth stages through non-invasive imaging.

As a first step, a reliable method for spike detection is needed. Once this is established, the time point of the first detection of spikes across a time course can be determined. To date, there have been relatively few studies concerned with wheat spike detection and growth analysis from digital images. Most of them are based on single spikes and needed to cut off spikes to classify different wheat varieties using morphological image processing algorithms, Hu moments, and neural networks (Kun et al., 2011; Bi et al., 2010, 2011). However, these methods are unsuitable for non-invasively detecting spikes from a whole plant with overlapping of leaves and young developing spikes in a high-throughput manner.

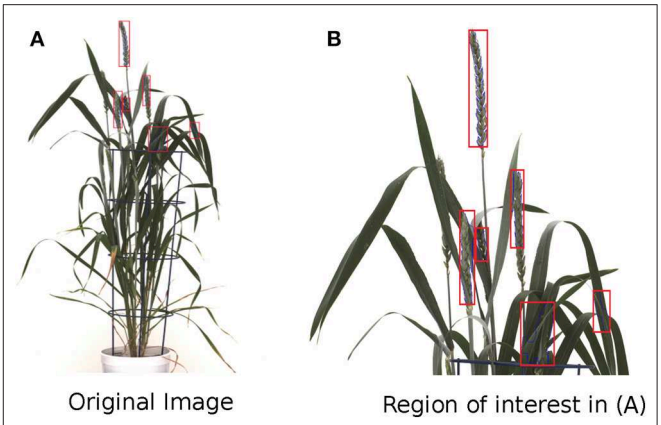

FIGURE 1 | Limitations of wheat spike detection using the Qiongyan et al. (2017) algorithm. **(A)** The detection of spike and non-spike pixels in the wheat plant. **(B)** Zoomed version of detected pixels in **(A)**.

Qiongyan et al. (2017) proposed a novel approach for detecting (young) spikes in digital images of wheat plants based on Law's textural (energy) features and a neural network. This approach is based on the fact that spikes and leaves have a high color similarity but differ clearly in texture. However, when we applied this algorithm to one of our data sets, it turned out to be sensitive to the high-energy leaf edges and tillers, which led to false classifications of spike and non-spike pixels (or noisy pixels) as shown in **Figure 1**. However, their method was based on four Australian wheat varieties. In contrast, our data set is based on a diverse collection of high-yielding mainly European elite cultivars that are much more diverse in their plant architecture and produce more leaves and biomass compared to Australian genotypes. Accordingly, due to the presence of noisy pixels in the final image segmentation, the heading time point (HTP) BBCH55 was detected too early on our dataset compared to the ground truth data using their method. Thus, solely depending on Law's textural features lead to false detection of spikes in our wheat panel. Therefore, to overcome these artifacts, an improved and extended novel approach is proposed in this paper.

The paper is structured as follows. Section 2 deals with the improved methodological framework of spike detection, including data preparation, segmentation, and post-processing algorithms. Section 3 describes the improvement of our algorithm compared to the existing method for HTP detection and the spike growth analysis. In summary (section 4), we draw conclusions regarding the performance of our algorithm and discuss its future improvements.

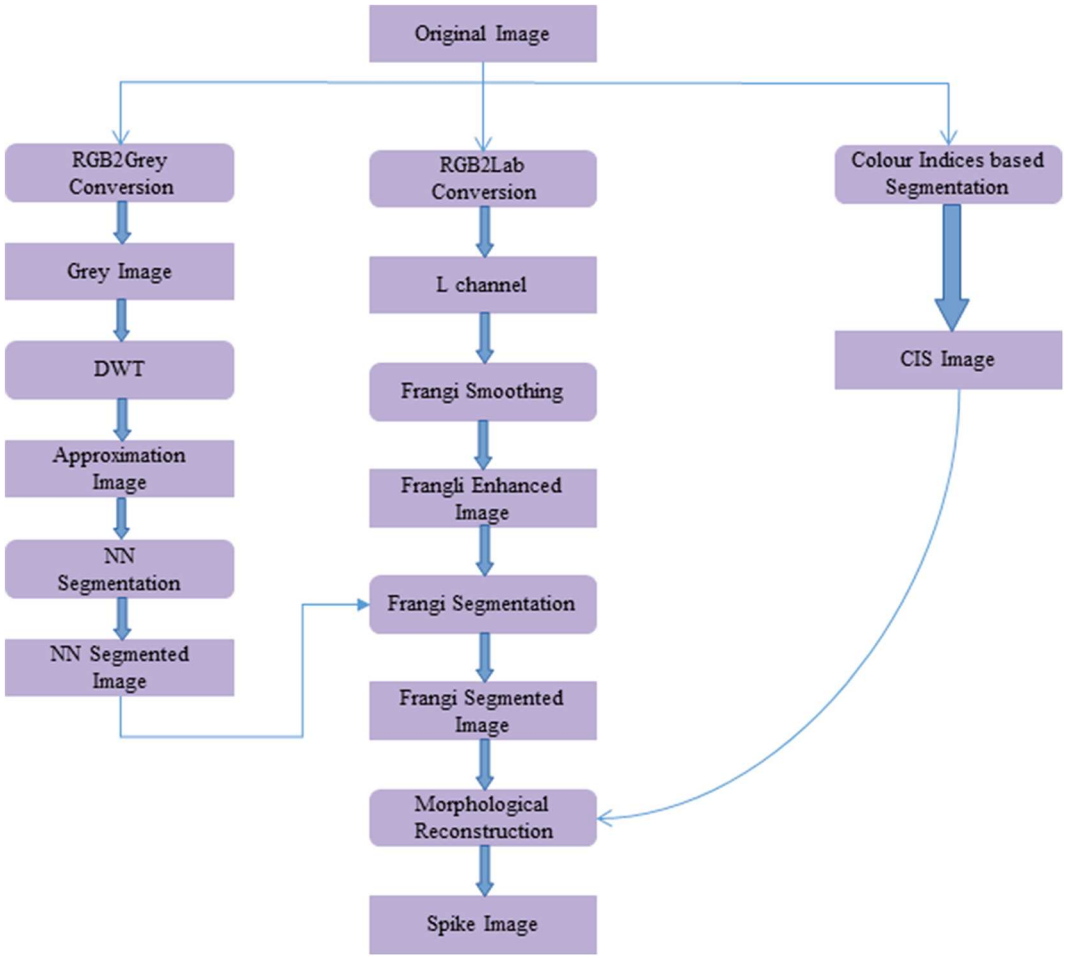

FIGURE 2 | Workflow of the proposed spike detection algorithm using image processing methods. Framed rectangles represent the data modalities, and other rectangles describe the image processing operations.

2. MATERIALS AND METHODS

2.1. Dataset

We used images from one experiment with 260 diverse winter wheat cultivars of mainly Central European origin. Of these lines, 220 correspond to the collection described in Voss-Fels et al. (2019) and represent high-yielding cultivars of the past decades. The remaining 40 lines are winter wheat elite cultivars from the Gabi-Wheat collection (Zanke et al., 2014), representing a similar breeding pool. Each cultivar was represented by two biological replicates. Sowing was done in small turf trays, and 14 days after sowing (DAS) at about the 2-leaf stage, plants were placed for vernalization into a growth chamber with an 8-h light period and 4°C day/night. After 8 weeks of vernalization, turf trays were placed in a greenhouse with 15-h light and 16°/12°C during the day/night for 3 days to acclimate the plants to higher temperatures. The plants were then repotted from the trays to 2-l-volume pots and were grown in the same greenhouse for another 7 days before they were placed on the imaging system, a LemnaTec 3D Scanalyzer (LemnaTec GmbH, Aachen, Germany). They were imaged and watered daily, with watering by target weight option corresponding to 89% of the plant-available water content in the soil (Dhanagond et al., 2019). Temperatures in the greenhouse of the imaging system were raised over the time course of the experiment from 16°/12°C in four steps to 30°/20°C to simulate a German spring/summer growing period, including 10 days of heat stress. In total, plants lasted 50 days on the imaging system before they were transferred to a normal greenhouse at 130 days after sowing (DAS) to grow to maturity and to evaluate the yield components. During the imaging period, the tiller number per plant was counted manually at the end of the heat period (at 125 DAS).

Images were taken from three side view angles (0°, 45°, and 90°) and one top view using RGB cameras. The top view camera (a Manta G-504) had a resolution of 2,452 × 2,056 pixels with a pixel size of 3.45 × 3.45 μm, while the side view camera had a resolution of 6,576 × 4,384 pixels and a pixel size of 5.5 × 5.5 μm. Plant images were later visually inspected to determine the time point of heading when the ear was half out of the flag leaf

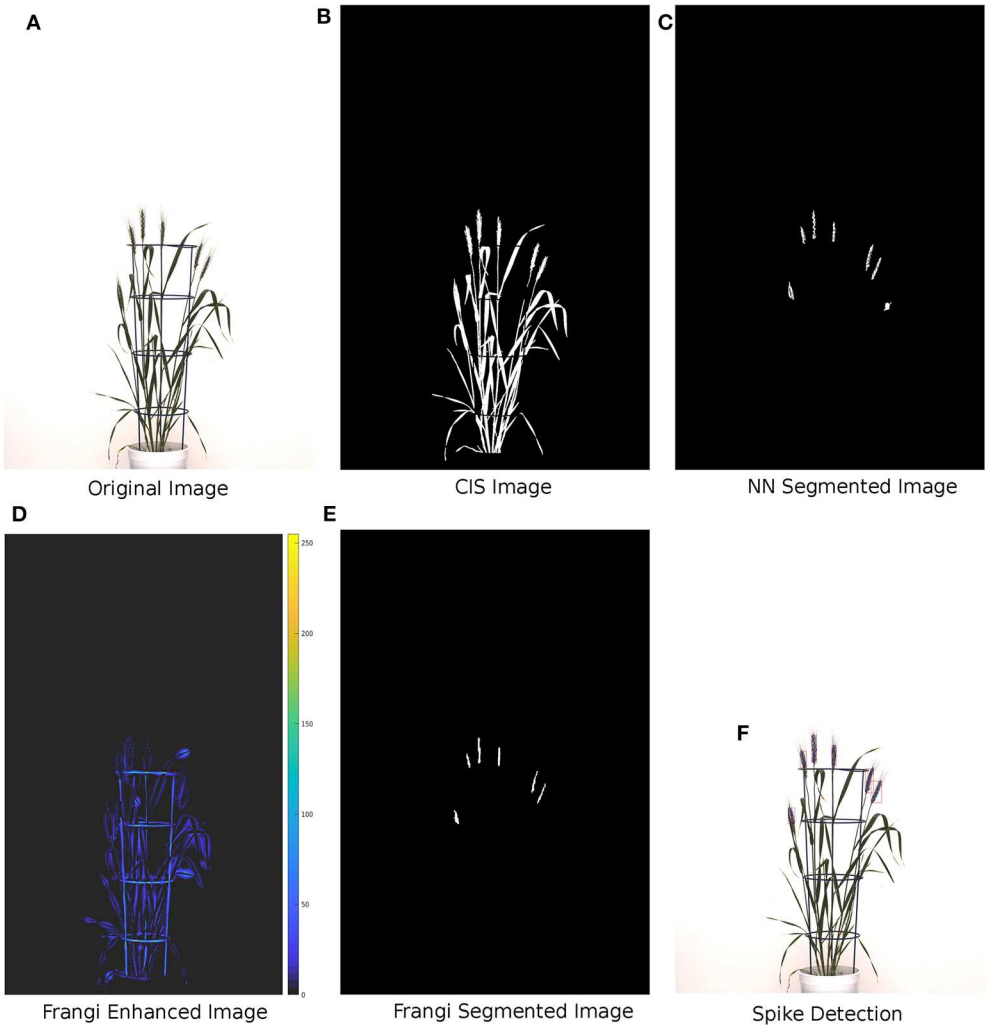

FIGURE 3 | Methodology of the proposed spike detection algorithm with example images. **(A)** Wheat plant with ID 1817KN373 at 150 days after sowing. **(B)** Green color indices-based segmented image. **(C)** DWT + Laws textural features-based NN segmented image. **(D)** Image Frangi enhanced at multiple scales and orientations. **(E)** Final binary segmentation: one leaf-crossing artifact is suppressed with the Frangi-enhanced image. **(F)** Spikes detected after morphological reconstruction.

(BBCH55). Here, top view images turned out not to be suitable as, from the top, an emerging ear has a very low visible area and might be easily hidden under a bending leaf. Moreover, it is hard to define how much of the ear is above the flag leaves. Therefore, this determination was done on inspecting the three side view images. In this case, only the pots were rotated; the camera is stable. Out of all 520 plants, 369 reached BBCH55 during the imaging period belonging to 202 different cultivars. These 369 plants from 202 genotypes were available for testing our spike detection algorithm. These plants exhibit strong differences in plant architecture and are challenging for this kind of analysis, for example, spikes with or without awns, short and tall plants (plant height range at harvest time from 34 to 119 cm), and especially low and high tillering genotypes ranging from 1 to 38 tillers per plant counted at 125 DAS during the imaging period. Further, the data set exhibits differences in BBCH55 timing of 29 days.

2.2. Methodology

The workflow for spike detection following image acquisition is shown in **Figure 2**. This algorithm was developed in the MATLAB environment (MATLAB 2019a). The methodology involved in the proposed algorithm is as follows:

In the initial step, the original image (**Figure 3A**) is converted to a grayscale image using MATLAB's *rgb2gray* routine. To enhance the separability between the plant and background pixels, discrete wavelet transform (DWT) is applied in the preprocessing step using the Haar basis function (Stanković and Falkowski, 2003). The DWT is a single level 2-D wavelet decomposition that produces a featured image called an approximation coefficients image (A). This image is half the size of the original image and is useful for characterizing unique textures. Then, a neural network-based Laws texture energy method is applied to image A, as proposed in Bi et al. (2010) and Qiongyan et al. (2017), to segment the spike pixels from

the plant pixels. Here, the segmentation of plant pixels from the background is called color index-based segmentation (CIS). Example images of the CIS and the neural network segmentation are shown in **Figures 3B,C**, respectively. However, the Laws texture energy is sensitive to the high-energy noisy edges (or pixels on leaves and leaf crossings) in the plant. To eliminate those noisy edges, a combination of a multi-scale Frangi-filtered image (Frangi et al., 1998) and the neural network segmented image is considered. Because the Frangi filter delivers a strength estimate of edges in the image, noisy edges can be suppressed by smoothing the image over multiple scales and orientations (Frangi et al., 1998). Therefore, this combination suppresses the tiny leaf edges and leaf crossings in the segmented image. Here, the Frangi filter is applied to an L component of the L*a*b color space image because the intensity values in the L component are closely matched with the human perception and contrast between the plant and non-plant pixels is high compared to in the a and b channels.

The Frangi-filtered image is considered one of the post-processing steps, because as a pre-processing step, it might lead to false representation of textures in the image. In other words, there might be a possibility of suppressing the spike pixels, hence modifying the unique textural characteristics of the spikes and leaves. Examples of a Frangi-filtered image and a segmented image are shown in **Figures 3D,E**, respectively. The complete spike is then recovered by applying morphological binary operations to the Frangi segmented image, as shown in **Figure 3F**.

2.2.1. Wavelet Decomposition

The wavelet-based texture classification is important because (1) it decorrelates the data (Fan, 2003) by stretching the color differences between plant and non-plant pixels in the image, and (2) it provides a non-redundant compressed image, which reduces the computation complexity significantly compared to the original grayscale image. Typically, wavelets are defined for 1-D signals, so extension to 2-D signals is usually performed by using a product of 1-D filters. The practical implementation of the wavelet transforms using different filters is as follows.

$$\begin{aligned} A &= [L_x * [L_y * I]_{\downarrow 2,1}]_{\downarrow 1,2} \\ H &= [L_x * [G_y * I]_{\downarrow 2,1}]_{\downarrow 1,2} \\ V &= [G_x * [L_y * I]_{\downarrow 2,1}]_{\downarrow 1,2} \\ D &= [G_x * [G_y * I]_{\downarrow 2,1}]_{\downarrow 1,2} \end{aligned} \tag{1}$$

where $*$ denotes the convolution operator, and $(\downarrow\ 2,1)$ and $(\downarrow\ 1,2)$ represent the downscaling along rows and columns, respectively. L and G are the low- and high-pass filters, and I is the original image. The DWT decomposes an image into four sub-bands called approximation coefficients (A), horizontal (H), vertical (V), and diagonal (D), as shown in **Figure 4**. Sub-band A is obtained by the low-pass filtering and is accordingly called the low-resolution image, the size of which is dependent on the level of decomposition and input image size. In contrast, H, V, and D are obtained by bandpass filtering in a specific direction. Therefore, they provide detailed directional information for the image. Among these sub-bands, A is an essential feature image (or coefficients image) bearing the textural information relevant to image segmentation. Consequently, the A wavelet coefficient image is used here for texture characterization.

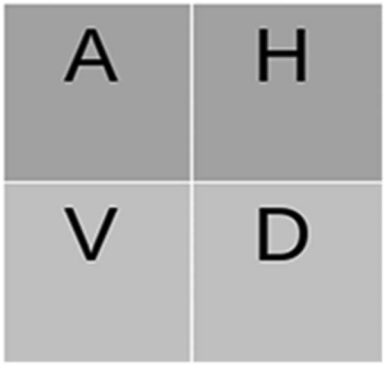

FIGURE 4 | DWT Decomposition: The coefficient image (A) is again decomposed in multilevel DWT decomposition.

2.2.2. Laws Textural Features-Based Image Segmentation Using Neural Networks

Laws' texture energy method (Laws, 1980) is a classical pixel-wise textural analysis approach and it has been used in many applications (Chang and Kuo, 1993; Jiang and Chen, 1998; Christodoulou et al., 2003; Mougiakakou et al., 2007). This approach uses 1-D local masks to detect various types of micro-structural textural features. The typical 1-D local masks are level detection, edge detection, and spot detection, as shown in Equation (2). However, the image is two-dimensional and requires 2-D masks for textural analysis.

$$\begin{aligned} L3 &= [1\ 2\ 1] \text{ - Level detection} \\ E3 &= [-1\ 0\ 1] \text{ - Edge Detection} \\ S3 &= [-1\ 2\ -1] \text{ - Spot Detection} \end{aligned} \tag{2}$$

The 2-D masks are generated from the 1-D masks by convolving the vertical 1-D mask with the horizontal 1-D mask. For example, mask S3L3 is calculated by convolving vertical mask S3 with horizontal mask L3 and is a zero-sum mask. In contrast, mask L3L3 is a non-zero-sum mask, which is not considered for the textural analysis. The list of 2-D masks used for the textural analysis is as follows:

$$\begin{aligned} L3E3 &= L3^T * E3;\ \ E3S3 = E3^T * E3; \\ L3S3 &= L3^T * S3;\ \ S3L3 = S3^T * L3; \\ E3L3 &= E3^T * L3;\ \ S3E3 = S3^T * E3; \\ E3E3 &= E3^T * E3;\ \ S3S3 = S3^T * S3; \end{aligned} \tag{3}$$

The textural features are calculated in two steps (Chang and Kuo, 1993) using 2-D masks. In the first step, the input image is convolved with all of the above 2-D masks. Then, each individual resulting image is normalized with a unit standard deviation and average mean over the window size of 25. Consequently, eight textural feature images are generated for every input image. However, these feature images have both plant and background

TABLE 1 | Statistical performance of the neural network in the training stage.

	Training	Testing	Validation	Total
Spike pixels	152793	32773	32716	218282
Non-spike pixels	511743	109627	109684	731054
TP rate (%)	96.2	96.4	96.0	96.2
TN rate (%)	99.3	99.3	99.3	99.3
Accuracy (%)	98.5	98.6	98.5	98.6

pixels, which increases the computational complexity of the neural network for spike detection.

To overcome the computational complexity of the image segmentation, the plant pixels (PP) are segmented from the background pixels using the CIS method (Bi et al., 2010) as follows.

$$PP = 2g - r - b \tag{4}$$

This method decorrelates the dominating green plant pixels from the background. A binary plant image is then generated using the binarization technique (pixel value > 0), see **Figure 3B**. As a result, the number of pixels for the neural network-based segmentation is reduced significantly.

The neural network is used to perform the classification of spike and non-spike pixels in the study. In practice, the neural network is trained with a large quantity of spike and non-spike pixels from the different wheat plants. The trained neural network parameters are then adapted to perform the spike detection in an automated manner. Here, a total of 218282 spike and 731054 non-spike pixels were extracted from 150 manually segmented images and subsequently used for training, testing, and validation of a network model in the sample proportion 70:15:15. The performance of the network model, with eight input nodes, one hidden layer with 10 hidden nodes, and 2 output nodes, was assessed using the conventional confusion matrix [TP FP; FN TN], components of which indicate the total number of correctly and incorrectly classified spike and non-spike pixels, respectively. The true positive (TP) and true negative (TN) rates, as well as the overall accuracy (TP+TN)/(TP+FP+FN+TN), are summarized in **Table 1**.

2.2.3. Frangi Filter Enhancement

The Frangi filter is a multi-scale second-order vessel enhancement method developed by Frangi et al. (1998) that is frequently used in biomedical applications (Vazquez et al., 2001; Budai et al., 2013; Shahid and Taj, 2018). The Frangi filter is used for enhancement of high-contrast vessel structures or edges along with the suppression of the non-vessel structures and thin vessel edges. Since wheat shoots have multiple leaf crossings, they exhibit vessel-like thin structures producing high-energy signals similar to spikes. In turn, this can lead to false spike detection at leaf crossings by the network model, as shown in **Figure 1**. The Frangi filter is applied to suppress edges resulting from such leaf crossings in the neural network segmented images.

Frangi-based vessel enhancement is achieved based on Hessian and eigenvalues. The Hessian matrix of image I is computed as follows:

TABLE 2 | Possible structural patterns in 2D images depending on eigenvalues λ_1 and λ_2.

λ_1	λ_2	Structure pattern
N	N	Noisy, no preferred direction
L	H−	Vessel structure (bright)
L	H+	Vessel structure (dark)
H−	H−	Blob like structure (bright)
H+	H+	Blob like structure (dark)

H = high, L = low, +/− indicates the sign of the eigenvalue (Frangi et al., 1998).

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \sigma \begin{bmatrix} \frac{\partial^2 I}{\partial_{x^2}} & \frac{\partial^2 I}{\partial_x \partial_y} \\ \frac{\partial^2 I}{\partial_y \partial_x} & \frac{\partial^2 I}{\partial_{y^2}} \end{bmatrix} \tag{5}$$

where $h_{11}, h_{12}, h_{21}, h_{22}$ are the second-order partial derivatives of the image and σ denotes a variable scaling factor.

To extract information on structural patterns from the Hessian matrix, the eigenvalues λ_1 and λ_2 are calculated, while σ is used for the enhancement of structures at different scales, see **Table 2**. Since we are interested in detecting and suppressing the bright vessel-like structures in the plant leaves, the image enhancement is performed under the assumption that a pixel belonging to a vessel region should have a very low value of λ_1 and a very high magnitude of λ_2; see Equation (6). Furthermore, the bright vessel-like structures emerge with negative λ_2, and the filter response of the corresponding pixel with $\lambda_2 > 0$ is considered to be zero in the enhanced image.

$$| \lambda_1 | \leq | \lambda_2 | \tag{6}$$

The enhanced image is defined as follows:

$$I_E = \begin{cases} \text{if } \lambda_2 > 0: & 0, \\ \text{otherwise}: & exp(\frac{-R_B^2}{2\beta^2})(1 - exp(\frac{-S^2}{2c^2})) \end{cases} \tag{7}$$

where $R_B = \frac{\lambda_1}{\lambda_2}$, $S = \sqrt{\lambda_1^2 + \lambda_2^2}$, and c, β are constants that control the sensitivity of the filter. The enhanced image I_E is obtained at various scales, i.e., $\sigma = 1, 3, 5, 7, 9$. Since the maximum scale approximately matches the size of the vessel to detect, the final enhanced image I_{FE} can be obtained according to Frangi et al. (1998) by taking a maximum among all scales as defined in Equation (8).

$$I_{FE} = \max_\sigma I_E \tag{8}$$

An example of edge suppression (leaf crossings) using the Frangi filter is shown in **Figure 5**.

Consequently, the result of the neural network segmentation is subsequently filtered under consideration of leaf-crossing regions detected by the Frangi filter (**Figure 3D**). This is done by eliminating the regions corresponding to leaf edges in the binary segmentation mask; see **Figure 3E**.

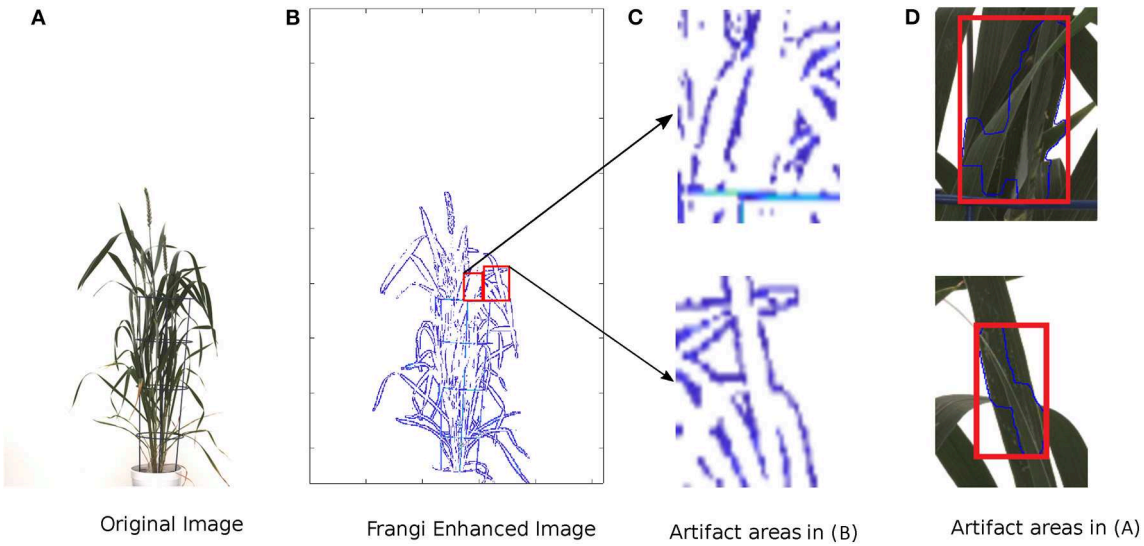

FIGURE 5 | Example of suppression of leaf crossings using the Frangi Filter. From left to right: **(A)** original image of a wheat shoot, **(B)** Frangi filter-enhanced image, **(C)** examples of Frangi-enhanced regions, **(D)** examples of leaf crossings detected in the original image.

2.2.4. Spike Reconstruction Using Morphological Filters

As shown in **Figure 3E**, only some parts of the spikes were detected using the proposed algorithm compared to the CIS image in **Figure 3B**. To recover the complete spikes, the logical "and" operation of the CIS image and the Frangi segmented image were performed. Then the morphological binary operations (erosion and dilation) were sequentially applied to recover the final spike area in the CIS image; see **Figure 3F**.

3. RESULTS AND DISCUSSION

The above-described algorithm was applied to calculate the yield-related features at the transition from the tillering to flowering growth stages of wheat plants with an age of more than 90 DAS. Accordingly, the results of this study are presented in two sections dedicated to (i) detection of the time point of spike emergence and (ii) spike growth analysis from RGB images acquired using visible light cameras during an experiment with diverse winter wheat varieties. In the first section, the spike emergence was tested on 369 wheat plants from 202 different genotypes. Here, the HTP was defined as the first time in the imaging time course when the detected spike satisfied the minimum area constraint of 500 pixels. The spike area was then measured in a time series from the HTP to perform real-time growth analysis for a few selected plants.

Image analysis was performed on an Intel Xeon CPU E5-2640-based workstation running under the Linux OS. The algorithms were implemented under MATLAB 2019a (MathWorks Inc.). Training of a neural network on 949336 manually segmented spike and non-spike pixels using ten 2.40GHz CPUs with a total of 20 cores in parallel mode took approximately 80 s. The spike detection algorithm takes approximately 2.5 s to process an 8-megapixel test image. However, the processing time might vary depending on the test image size.

The root mean square error (RMSE) is used for quantification of the deviations of predictions from our model and Qiongyan et al.'s model from ground truth data,:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

where y is the ground truth value and $\hat{y}$ is the model-predicted value. For consistent comparison of performance, the Qiongyan et al. model was retrained with the European cultivars.

3.1. Spike Emergence

The time-series images of a single plant described in Section 2 have three orientations. Accordingly, two factors are considered to estimate the HTP from multiple orientations: the spike should (1) appear in at least two orientations and (2) remain present in all days of the experiment. This means the spike or spikes should be continuously detected until the last day to avoid false emerging time points.

Figure 6 shows HTP detection in the wheat plant side-view images. These nine different wheat plants from the early-flowering genotypes possessing a single spike (1817KN397, 1817KN422) and multiple spikes (remaining seven plants) were considered for the training a model because we were aware that the later-flowering genotypes, which produce more biomass, will present much greater difficulties with spike visibility due to a higher probability that the spike will be covered by leaves. **Figure 6** indicates that HTP values obtained by the proposed method have a significant correlation with the ground-truth HTPs, with an RMSE of 1.94, whereas the Qiongyan et al. method

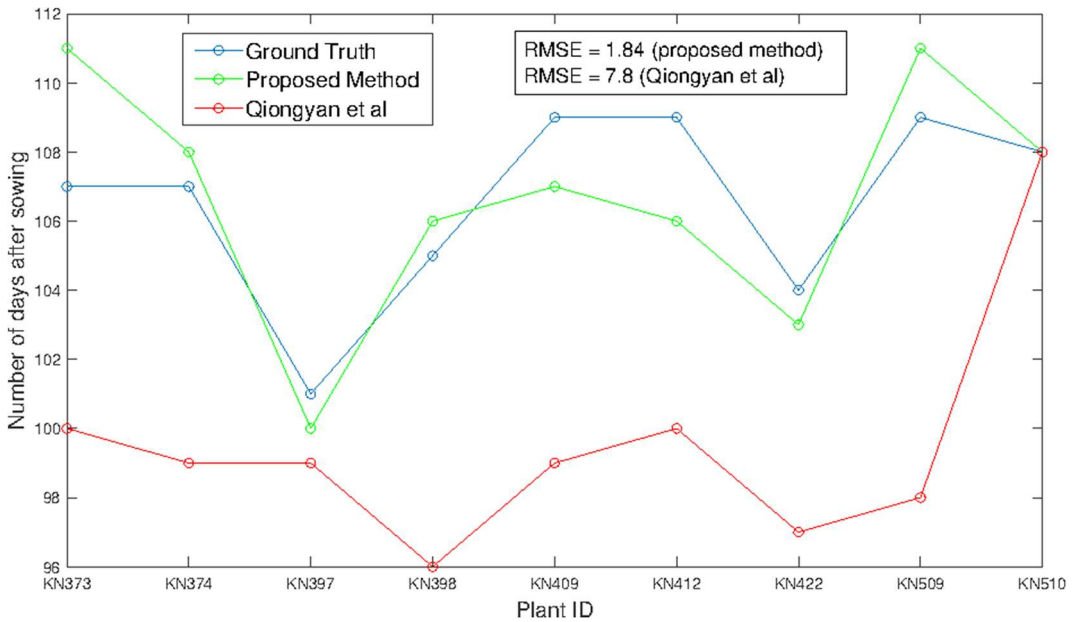

FIGURE 6 | HTP detection using the method of Qiongyan et al. (2017) and our proposed method in comparison to the ground truth.

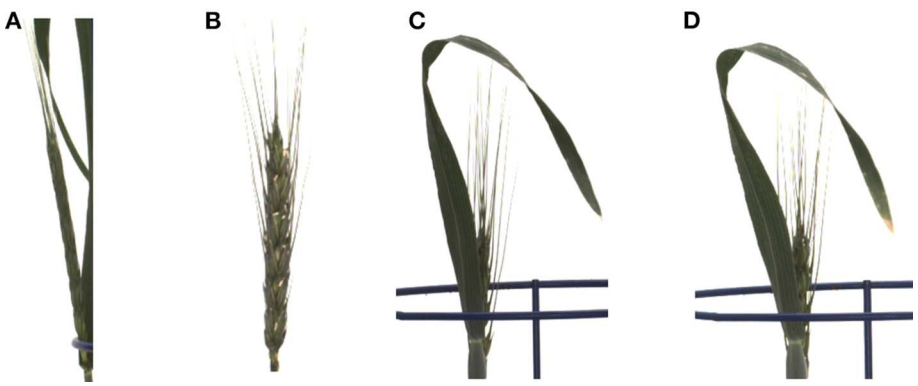

FIGURE 7 | Limitations of the proposed method. **(A)** The early-stage spike texture failed to be detected in plant ID 1817KN373. **(B)** The detected spike texture in plant ID 1817KN373. **(C)** Example spike geometry less than the BBCH55 scale in plant ID 1817KN412. **(D)** Spike geometry according to the BBCH scale in plant ID 1817KN412.

underperforms, with an RMSE of 7.8. This indicates that the Qiongyan et al. method is highly sensitive to the leaf artifacts whose energy is similar to that of the spike pixels but that those leaf artifacts were suppressed by the proposed method, as shown in **Figure 5**.

On the other hand, the proposed method resulted in high HTP error rates of 4 days more and 3 days less for plant ID 1817KN373 and 1817KN412, respectively. For plant ID 1817KN373, this was because the spikes were narrow and the pixel-wise textural energy was similar to that of the leaves, as shown in **Figure 7A** compared to the other spikes in **Figure 7B**. Therefore, the HTP was detected 4 days later. In the case of plant ID 1817KN412, it turned out that the visually scored time point was determined too late, most likely by not carefully inspecting all side view angles (in the first, at 0°, the later time point looks correct, but at the 45° and 90° angles, it is visible that the earlier one is correct). Example spike images for the early HTP detection are shown in **Figures 7C,D**.

The advantages and significance of the results with the proposed method showed that it is feasible for high-throughput analysis of HTP detection. Consequently, we applied the method to all 369 diverse plants in our data set that reached heading within the imaging period. As expected, 104 plants corresponding to the supposedly earlier-heading genotypes obtained a good and reliable estimation of the true heading time point. **Figure 8** shows the results for the high-throughput analysis of 104 plants. It is observed that the proposed method outperforms the Qiongyan et al. method, with an R^2 value of 0.776 compared to the R^2 value (0.193) of the Qiongyan et al. method. Since the European elite cultivars possess more leaves, overlay artifacts result in too early HTP detection using the Qiongyan et al.

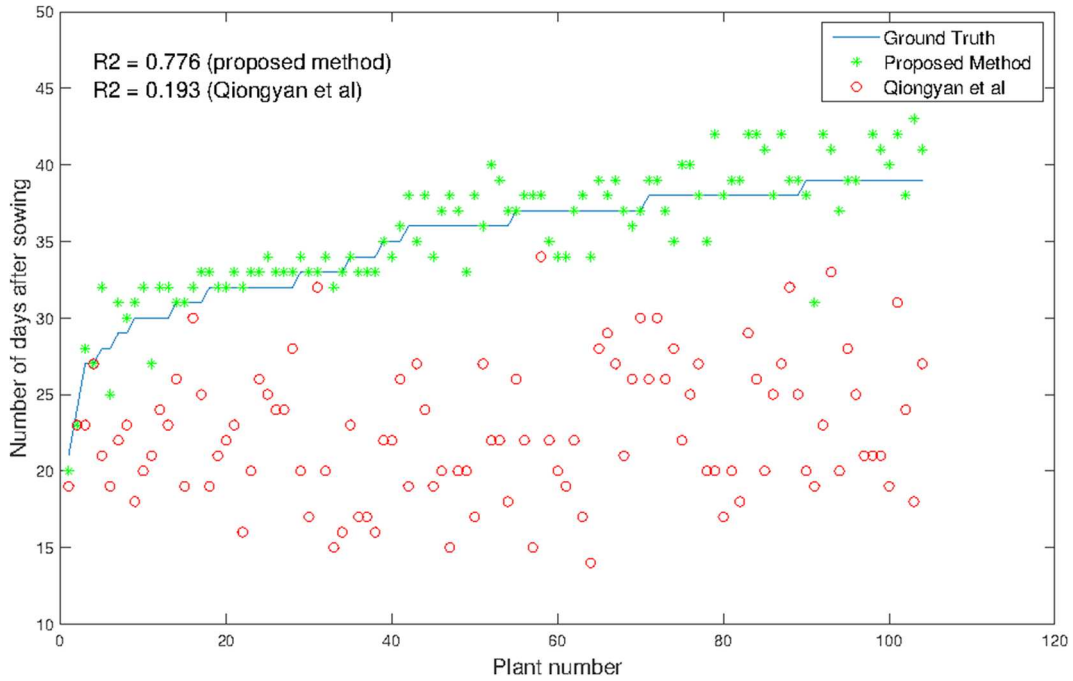

FIGURE 8 | Comparison of HTP detection using our method and that of Qiongyan et al. vs. ground truth in 104 wheat plants.

TABLE 3 | Gene classification and comparison of architectural features of 369 plants.

Phenotypic traits	Phenotypic trait mean values		
	2 out of 2 plants successful	1 plant of 2 successful	0 out of 2 plants successful
Ground truth BBCH55 (DAS)	115.5 (107–120)	118.1 (101–127)	125.5 (120–130)
Days to maturity (DAS)	175.4 (159–203)	185.2 (160–222)	193.8 (166–283)
Presence of awns (1=no, 2=yes)	1.3	1.2	1.0
Final plant height (cm)	57.1 (34–101)	64.0 (37–96)	60.9 (38–119)
Tiller number at DAS 125	7.5 (3–19)	8.4 (1–17)	11.4 (4–38)
Spike number at harvest	7.5 (3–16)	7.8 (1–14)	9.8 (4–22)
Total plant biomass at harvest (grains + straw) (g)	15.2 (5.7–26.8)	17.5 (4.5–28.1)	21.4 (8.1–48.0)
Total plant straw weight at harvest (g)	9.9 (3.5–15.6)	12.8 (5.7–20.0)	15.7 (5.8–38.2)

method on 90% of our data. In the remaining 265 plants, the spike emerged in the final days according to the ground truth data, and they have early-stage spike textural features that are similar to the leaves. This resulted in the proposed algorithm failing to detect the spikes in the final days with a day number 0 in the output. This leads to a low value of the correlation coefficient R^2 0.0586 for the remaining 265 plants.

We compared the general plant architecture features of all 369 plants tested and classified them into three categories: (i) both plants of the genotype were classified correctly by our algorithm (94 plants from 47 genotypes, (ii) only one out the two plants of a genotype were classified correctly by our algorithm (20 plants from 10 genotypes), and (iii) none of the two plants of a genotype were classified correctly by our algorithm (**Table 3**; **Supplementary Material**) . It turned out that the method performed better for earlier-flowering plants with an accordingly lower number of tillers and less biomass. Moreover, in 26 out of all 39 plants with awned spikes, heading time could be reliably estimated by our algorithm. This might arise from two factors: first, awned genotypes are more abundant in the earlier-flowering group and possess less biomass, and therefore spikes are less often hidden by leaves, and second, the model was trained based on nine early-flowering plant IDs with a bias toward awned types. Further, it might very well be that the fine awn structures, in general, help in the differentiation of the spike from the leaf background.

Table 3 shows mean phenotypic trait values, with minimum and maximum in brackets, of plants successfully and non-successfully classified regarding their time point of heading.

3.2. Spike Area

Spike area is one of the key yield measures in wheat plants, so we have examined the total spike growth of a single wheat plant in three orientations from the spike emergence day for all images. In section 3.1, nine wheat plants were considered for HTP detection. Among those, only three plants (1817KN374, 1817KN409, and 1817KN422) with a single spike and two with multiple spikes are considered for the spike growth analysis. Here, the spike area of a

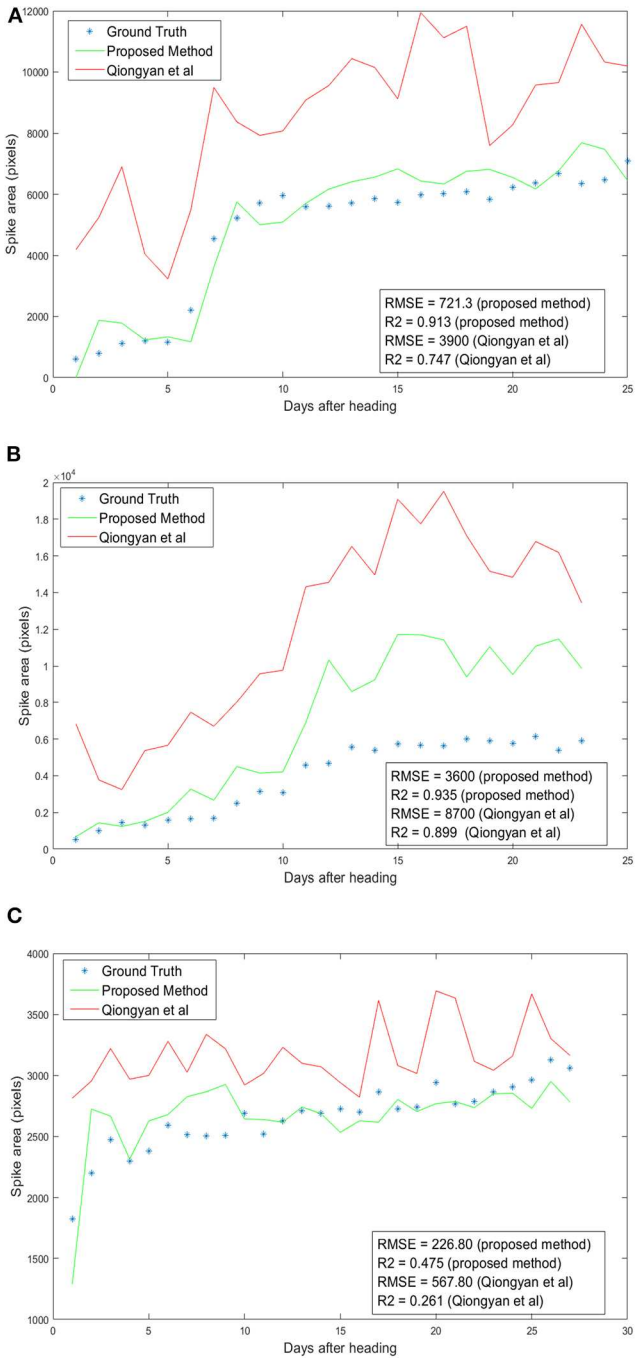

FIGURE 9 | Spike growth analysis: Day number 1 represents the first day of ear emergence in the wheat plants. **(A)** Plant ID 1817KN374 with multiple spikes. **(B)** Plant ID 1817KN409 with multiple spikes. **(C)** Plant ID 1817KN422 with a single spike.

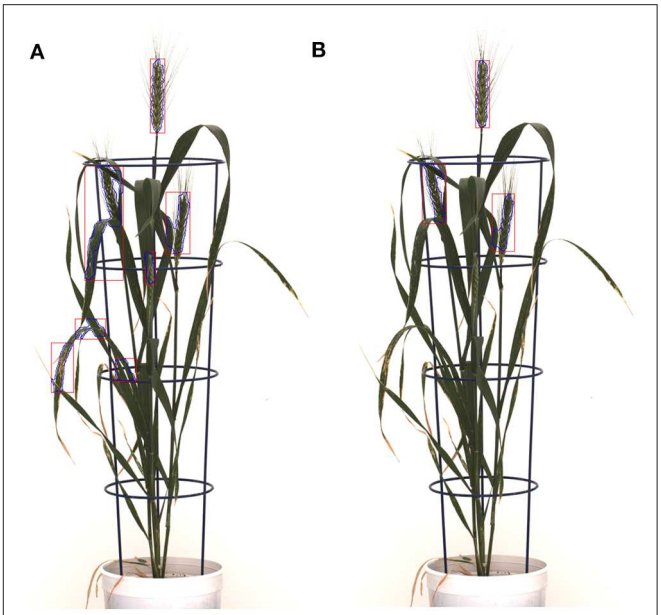

FIGURE 10 | The detected leaf artifacts in **(A)** result in a high spike area compared to **(B)** for the spike growth analysis. The segmented objects are represented with blue curves and red rectangular boxes; **(A)** the Qiongyan et al. method and **(B)** the proposed method.

plant per day is calculated by taking the maximum area among the three orientations. The measured area of both algorithms is validated by the RMSE and R^2. The RMSE quantifies the difference between the ground truth area and the predicted area for all days from the ear-emergence day. The R^2 value compares the goodness of our proposed models and of the Qiongyan et al. model compared to the ground truth data.

Figure 9 shows the results of spike growth analysis with the Qiongyan et al. method and with our proposed method compared to the ground truth data. Here, the ground truth data are prepared manually by segmenting the spikes using GIMP image processing software (https://www.gimp.org). The number of non-zero pixels in the segmented image represents the actual spike area or the ground-truth spike area of the image. This figure shows that the proposed method outperforms the Qiongyan et al. method overall, with a low RMSE and a high value of R^2. Moreover, the RMSE is profoundly improved by more than 50% and the R^2 value is significantly improved for plant ID 1817KN373 (**Figure 9A**) and plant ID 1817KN422 (**Figure 9C**). Nevertheless, plant ID 1817KN409 (**Figure 9B**) exhibits a high RMSE compared to the other plants in the spike growth analysis.

The high RMSE value for the Qiongyan et al. method is caused by the classification of leaf artifacts as spikes, which leads to an increase in the total spike area. In our method, these artifacts were eliminated using DWT and the Frangi filter. Example images of the improved spike detection are shown in **Figure 10**. On the other hand, the high error rate observed for plant ID 1817KN409 is due to the morphological reconstruction at the final step. This leads to the fusion of neighboring spikes with the connected stems and leaves, as shown in **Figure 11**.

4. CONCLUSION

Here, we present an improved method for wheat spike detection in a test data set with 369 plants from 202 diverse winter wheat varieties corresponding to mainly high-yielding Central

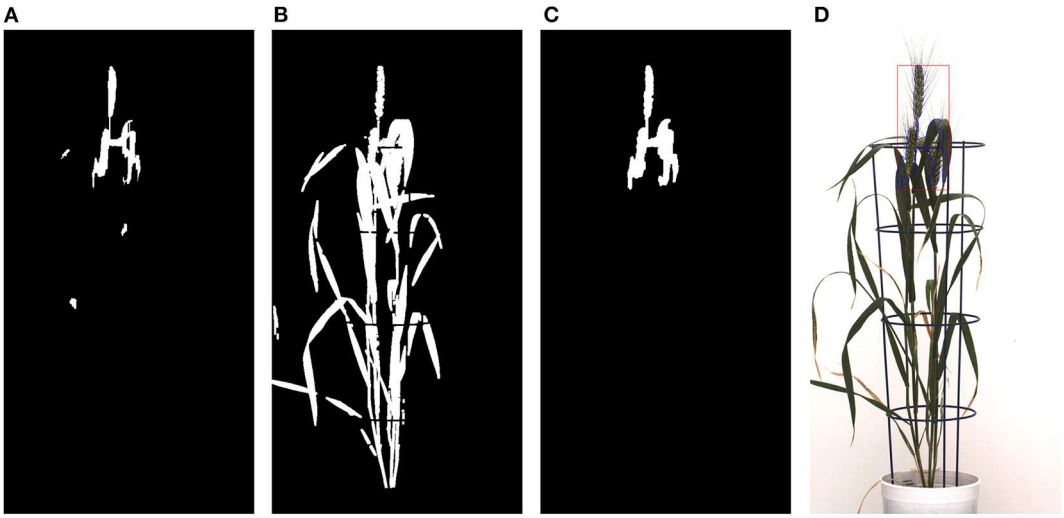

FIGURE 11 | Morphological reconstruction of the spikes: **(A)** Frangi-based spike segmentation. **(B)** CIS image. **(C)** Morphologically reconstructed image using **(A,B)**. **(D)** Spikes detected in the original image represented with blue lines and a red rectangular box.

European varieties (Voss-Fels et al., 2019). Our work relies on the algorithm proposed by Qiongyan et al. (2017), which was originally tailored to four Australian wheat varieties. By application to European elite cultivars, that earlier algorithm turned out to be too sensitive to the leaf crossing or overlay artifacts and aged leaves. This resulted in a high rate of false detection of spikes and, consequently, incorrect (too early) detection of heading time points. To overcome these limitations, we developed and evaluated an algorithmic pipeline extended by DWT and the Frangi filter that enable detection and suppression of high-energy regions caused by a high density of leaves. The proposed method has significantly improved the accuracy of the detection of spikes and the time point of heading, resulting in a reduction of the error rate (RMSE) by 75% compared to the Qiongyan et al. model. Similar improvement was also achieved in the analysis of spike growth, where the error rate of model predictions vs. ground truth data was reduced by 50% compared to Qiongyan et al. With these algorithmic improvements, detection of the heading time and analysis of spike growth can be performed in a high-throughput manner with sufficiently high accuracy.

In contrast to the majority of previous method studies, our approach was tested on a diverse set of genotypes with strong morphological differences regarding spike architecture (with or without awns), height, tiller number, biomass, and heading time. Such a data set is very challenging as it is easier to find an algorithm for identifying the plant organs in a small genotype set with much more uniform morphology. However, the biological truth is that many studies employ non-invasive phenotyping to screen genotype collections that exhibit a high phenotypic diversity (Honsdorf et al., 2014; Dhanagond et al., 2019). This requires algorithms with high performance across a highly heterogeneous background. Our proposed method represents a good starting point, as it correctly determined the heading date in 47 genotypes for both biological replicates and for at least one of the two biological replicates in a further 10 genotypes. These were mainly plants from lower biomass and tillering range and, therefore, on-average earlier heading. The method is thus expected to perform well in germplasm with relatively low biomass and tillering, as would be the case for collections from hot or dry environments. However, it also clearly showed limitations in genotypes with high biomass and high tillering (mostly later-heading types), where the spike emerges within a green "bush." The fine-tuning or extension of the developed method for reliable spike detection in such high-biomass, high-tillering genotypes will be conducted in the near future. Further, we aim at application to other existing data sets of spring barley and spring wheat collections, where ground truth data still have to be generated. It is likely that in collections with many or exclusively awned genotypes, the method would already be applicable and yield meaningful results. It is also conceivable that the presented method will work well in bi-parental mapping populations if both parents come from the lower-biomass and tiller-number spectrum.

In conclusion, the proposed approach has the potential to predict the spike yield in other cereal plants such as barley, rice, and rye over time.

In the future, we shall explore the possibility of advancing spike detection methods in an automated manner using deep learning technologies. We also plan to perform a time series analysis of spike growth over a large experimental population ($>$ 500 plants) to further improve the algorithm and to deliver more sophisticated solutions for plant breeders and cereal crop researchers.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

NN and EG conceived, designed, and performed the computational experiments, analyzed the data, wrote the paper, prepared figures and tables, and reviewed drafts of the paper. KN and MR executed the laboratory experiments, acquired image data, wrote, and reviewed drafts of the paper.

FUNDING

This work was performed within the German Plant-Phenotyping Network (DPPN), which is funded by the German Federal Ministry of Education and Research (BMBF) (project identification number: 031A053).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00666/full#supplementary-material

REFERENCES

Abdel-Ghani, A. H., Sharma, R., Wabila, C., Dhanagond, S., Owais, S. J., Duwayri, M. A., et al. (2019). Genome-wide association mapping in a diverse spring barley collection reveals the presence of qtl hotspots and candidate genes for root and shoot architecture traits at seedling stage. *BMC Plant Biol.* 19:216. doi: 10.1186/s12870-019-1828-5

Bi, K., Jiang, P., Li, L., Shi, B., and Wang, C. (2010). Non-destructive measurement of wheat spike characteristics based on morphological image processing. *Trans. Chin. Soc. Agric. Eng.* 2010, 212–216. doi: 10.3969/j.issn.1002-6819.2010.12.036

Bi, K., Huang, F.-F., and Wang, C. (2011). "Quick acquisition of wheat ear morphology parameter based on imaging processing," in *Computer Science for Environmental Engineering and EcoInformatics*, eds Y. Yu, Z. Yu, and J. Zhao (Berlin; Heidelberg. Springer), 300–307. doi: 10.1007/978-3-642-22694-6_42

Braun, H.-J., Atlin, G., and Payne, T. (2010). Multi-location testing as a tool to identify plant response to global climate change. *Clim. Change Crop Product.* 1, 115–138. doi: 10.1079/9781845936334.0115

Budai, A., Bock, R., Maier, A., Hornegger, J., and Michelson, G. (2013). Robust vessel segmentation in fundus images. *Int. J. Biomed. Imaging*, 2013, 1–11. doi: 10.1155/2013/154860

Chang, T., and Kuo, C.-C. J. (1993). Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. Image Process.* 2, 429–441. doi: 10.1109/83.242353

Christodoulou, C. I., Pattichis, C. S., Pantziaris, M., and Nicolaides, A. (2003). Texture-based classification of atherosclerotic carotid plaques. *IEEE Trans. Med. Imaging* 22, 902–912. doi: 10.1109/TMI.2003.815066

Cockram, J., Jones, H., Leigh, F. J., O'Sullivan, D., Powell, W., Laurie, D. A., et al. (2007). Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. *J. Exp. Bot.* 58, 1231–1244. doi: 10.1093/jxb/erm042

Dhanagond, S., Liu, G., Zhao, Y., Chen, D., Grieco, M., Reif, J., et al. (2019). Non-invasive phenotyping reveals genomic regions involved in pre-anthesis drought tolerance and recovery in spring barley. *Front. Plant Sci.* 10:1307. doi: 10.3389/fpls.2019.01307

Digel, B., Tavakol, E., Verderio, G., Tondelli, A., Xu, X., Cattivelli, L., et al. (2016). Photoperiod-H1 (PPD-H1) controls leaf size. *Plant Physiol.* 172, 405–415. doi: 10.1104/pp.16.00977

Fan, Y. (2003). On the approximate decorrelation property of the discrete wavelet transform for fractionally differenced processes. *IEEE Trans. Inform. Theory* 49, 516–521. doi: 10.1109/TIT.2002.807309

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI'98*, eds W. M. Wells, A. Colchester, and S. Delp (Berlin; Heidelberg: Springer), 130–137. doi: 10.1007/BFb0056195

Guo, Z., Liu, G., Röder, M. S., Reif, J. C., Ganal, M. W., and Schnurbusch, T. (2018). Genome-wide association analyses of plant growth traits during the stem elongation phase in wheat. *Plant Biotechnol. J.* 16, 2042–2052. doi: 10.1111/pbi.12937

Habte, E., Müller, L. M., Shtaya, M., Davis, S. J., and von Korff, M. (2014). Osmotic stress at the barley root affects expression of circadian clock genes in the shoot. *Plant Cell Environ.* 37, 1321–1337. doi: 10.1111/pce.12242

Honsdorf, N., March, T. J., Berger, B., Tester, M., and Pillen, K. (2014). High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. *PLoS ONE* 9:e97047. doi: 10.1371/journal.pone.0097047

Jiang, C.-F., and Chen, M.-L. (1998). "Segmentation of ultrasonic ovarian images by texture features," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)* (IEEE), 850–853. doi: 10.1109/IEMBS.1998.745570

Kun, B., Pan, J., Chongwei, T., Feifei, H., and Cheng, W. (2011). The design of wheat variety bp classifier based on wheat ear feature. *Chin. Agric. Sci. Bull.* 6, 465–467.

Laws, K. I. (1980). *Textured Image Segmentation*. Technical report, University of Southern California Los Angeles Image Processing INST. doi: 10.21236/ADA083283

Maeoka, R. E., Sadras, V. O., Ciampitti, I. A., Diaz, D. R., Fritz, A. K., and Lollato, R. P. (2020). Changes in the phenotype of winter wheat varieties released between 1920 and 2016 in response to in-furrow fertilizer: Biomass allocation, yield, and grain protein concentration. *Front. Plant Sci.* 10:1786. doi: 10.3389/fpls.2019.01786

Mougiakakou, S. G., Golemati, S., Gousias, I., Nicolaides, A. N., and Nikita, K. S. (2007). Computer-aided diagnosis of carotid atherosclerosis based on ultrasound image statistics, laws' texture and neural networks. *Ultrasound Med. Biol.* 33, 26–36. doi: 10.1016/j.ultrasmedbio.2006.07.032

Neumann, K., Zhao, Y., Chu, J., Keilwagen, J., Reif, J., Kilian, B., et al. (2017). Genetic architecture and temporal patterns of biomass accumulation in spring barley revealed by image analysis. *BMC Plant Biol.* 17:137. doi: 10.1186/s12870-017-1085-4

Qiongyan, L., Cai, J., Berger, B., Okamoto, M., and Miklavcic, S. J. (2017). Detecting spikes of wheat plants using neural networks with laws texture energy. *Plant Methods* 13:83. doi: 10.1186/s13007-017-0231-1

Shahid, M., and Taj, I. A. (2018). Robust retinal vessel segmentation using vessel's location map and Frangi enhancement filter. *IET Image Process.* 12, 494–501. doi: 10.1049/iet-ipr.2017.0457

Snape, J., Butterworth, K., Whitechurch, E., and Worland, A. J. (2001). *Waiting for Fine Times: Genetics of Flowering Time in Wheat*. Dordrecht: Springer. doi: 10.1007/978-94-017-3674-9_7

Stanković, R. S., and Falkowski, B. J. (2003). The Haar wavelet transform: its status and achievements. *Comput. Electric. Eng.* 29, 25–44. doi: 10.1016/S0045-7906(01)00011-8

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science* 327, 818–822. doi: 10.1126/science.1183700

Vazquez, M., Huyhn, N., and Chang, J. (2001). *Multi-scale vessel extraction using curvilinear filter-matching applied to digital photographs of human placentas* (Ph.D. thesis), California State University, Long Beach, CA.

Voss-Fels, K., Stahl, A., Wittkop, B., Lichthardt, C., Nagler, S., Rose, T., et al. (2019). Breeding improves wheat productivity under contrasting agrochemical input levels. *Nat. Plants* 5, 706–714. doi: 10.1038/s41477-019-0445-5

Witzenberger, A., and Hack, H. (1989). *Explanations of the BBCH Decimal Code for the Growth Stages of Cereals-With Illustrations*. Gesunde Pflanzen.

Zanke, C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al. (2014). Whole genome association mapping of plant height in winter wheat (*Triticum aestivum* L.). *PLoS ONE* 9:e113287. doi: 10.1371/journal.pone.0113287

Zhang, K., Tian, J., Zhao, L., Liu, B., and Chen, G. (2009). Detection of quantitative trait loci for heading date based on the doubled haploid progeny of two elite Chinese wheat cultivars. *Genetica* 135, 257–265. doi: 10.1007/s10709-008-9274-6

Zhao, Z., Rebetzke, G. J., Zheng, B., Chapman, S. C., and Wang, E. (2019). Modelling impact of early vigour on wheat yield in dryland regions. *J. Exp. Bot.* 70, 2535–2548. doi: 10.1093/jxb/erz069

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 Plant Science

METHODS
published: 09 July 2020
doi: 10.3389/fpls.2020.01040

Comparison Between Core Set Selection Methods Using Different Illumina Marker Platforms: A Case Study of Assessment of Diversity in Wheat

*Behnaz Soleimani[1], Heike Lehnert[2], Jens Keilwagen[2], Joerg Plieske[3], Frank Ordon[1], Sara Naseri Rad[4], Martin Ganal[3], Sebastian Beier[5] and Dragan Perovic[1]**

[1] Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Julius Kuehn Institute, Quedlinburg, Germany, [2] Institute for Biosafety in Plant Biotechnology, Julius Kuehn Institute, Quedlinburg, Germany, [3] TraitGenetics GmbH, Gatersleben, Germany, [4] Department of Physiology and Cell Biology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [5] Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany

OPEN ACCESS

Edited by:
Seth Jon Davis, University of York, United Kingdom

Reviewed by:
Amanda J. Burridge, University of Bristol, United Kingdom
Ravi Valluru, University of Lincoln, United Kingdom

***Correspondence:**
Dragan Perovic
dragan.perovic@julius-kuehn.de

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *31 January 2020*
Accepted: *24 June 2020*
Published: *09 July 2020*

Citation:
Soleimani B, Lehnert H, Keilwagen J, Plieske J, Ordon F, Naseri Rad S, Ganal M, Beier S and Perovic D (2020) Comparison Between Core Set Selection Methods Using Different Illumina Marker Platforms: A Case Study of Assessment of Diversity in Wheat. Front. Plant Sci. 11:1040. doi: 10.3389/fpls.2020.01040

Collections of plant genetic resources stored in genebanks are an important source of genetic diversity for improvement in plant breeding programs and for conservation of natural variation. The establishment of reduced representative collections from a large set of genotypes is a valuable tool that provides cost-effective access to the diversity present in the whole set. Software like Core Hunter 3 is available to generate high quality core sets. In addition, general clustering approaches, *e.g.*, *k*-medoids, are available to subdivide a large data set into small groups with maximum genetic diversity between groups.

Illumina genotyping platforms are a very efficient tool for the assessment of genetic diversity of plant genetic resources. The accumulation of genotyping data over time using commercial genotyping platforms raises the question of how such huge amount of information can be efficiently used for creating core collections. In the present study, after developing a 15K wheat Infinium array with 12,908 SNPs and genotyping a set of 479 hexaploid winter wheat lines (*Triticum aestivum*), a larger data set was created by merging 411 lines previously genotyped with the 90K iSelect array. Overlaying the markers from the 15K and 90K arrays enabled the identification of a common set of 12,806 markers, suggesting that the 15K array is a valuable and cost-effective resource for plant breeding programs.

Finally, we selected genetically diverse core sets out of these 890 wheat genotypes derived from five collections based on the common markers from the 15K and 90K SNP arrays. Two different approaches, *k*-medoids and Core Hunter 3 were compared,and *k*-medoids was identified as an efficient method for selecting small core sets out of a large collection of genotypes while retaining the genetic diversity of the original population.

Keywords: molecular marker, *Triticum aestivum*, k-medoids, core set, 90K–15K-iSelect Illumina arrays, SNP

INTRODUCTION

Germplasm collections are an important source of natural genetic diversity and provide a source of novel traits for sustainable crop improvement (Wang et al., 2018). However, genebanks need to balance between storing and regenerating large collections with limited resources with respect to storage capacity and monetary constraints. Frankel (1984) introduced the term core collection as a concept. A core collection is a subset of accessions which were selected by eliminating closely related samples while still capturing the genetic diversity of the original set of accessions. Therefore, a core collection ideally represents the genetic diversity of the entire collection. Providing core collections with maximum genetic variation facilitates efficient management and utilization of genetic diversity (Brown, 1989; van Heerwaarden et al., 2013) and is an efficient method for characterizing and using genetic resources of crop plants without the need to sample the entire collection (Jeong et al., 2017). Originally, phenotypic data containing both morphological and agronomical traits were used to create core collections, whereas nowadays molecular markers as neutral tools for measuring genetic variation have become the tool of choice.

There are currently three different strategies for generating a core collection from a large population using molecular marker data (Odong et al., 2013). Firstly, it is possible to build up a core collection that represents the individual accessions (CC-I), *e.g.*, a uniform representation of the original population. Second, it is possible to select a core collection based on accessions that represents the distribution of all relevant traits (CC-D), *e.g.*, if the majority of the original population contains allele A at a given locus, then the core collection should imitate this behavior. Thirdly, accessions can be selected that represent the extremes of all relevant traits (CC-X), *e.g.*, different entries into the core collection should be as diverse as possible with regard to the selected traits. Depending on which strategy is used, there are disadvantages in terms of working with the whole population. For example, trait customized core collections (CC-X), which aim to maximize diversity for that particular trait, would be better suited to finding rare alleles than a core collection that is designed to represent the original population (CC-I). The loss of rare alleles, especially in plant and resistance breeding, is one of the main concerns when working with core collections (Odong et al., 2013).

The quality of core set selection can be evaluated by using a variety of mathematical measures. De Beukelaer et al. (2018) explained that distance-based measures are attractive because they are easy to understand and take into account both the diversity within the core set and representativeness of accessions from the entire collection. Nevertheless, pairwise distances are required to be aggregated in suitable ways to evaluate the quality of a selected core set. One such aggregation, which is often used, is to calculate the average of pairwise distances to obtain an estimate of the quality of the core set (De Beukelaer et al., 2018). The interpretation of the result depends strongly on the defined purpose of the core collection. While it might be advantageous for core collections built up with the aim of conserving extremely rare alleles (CC-X) and therefore aiming at a maximum of the average pairwise genetic distance, a core collection built up for a uniform representation of the population (CC-I) would want to minimize the average pairwise genetic distance. Odong et al. (2013) proposed different criteria for the evaluation of core collections. They defined a way to estimate the quality of the core set selection process and introduced two new distance-based metrics. These two metrics were also used in the study by De Beukelaer et al. (2018) to evaluate the quality of core collections in rice, coconut, maize and pea for various tools. Core Hunter 3 was able to convince particularly through its flexibility to combine different methods. The evaluation metrics used showed that Core Hunter 3 core collections were always competitive with other more specialized methods.

An increasing number of plant genetic resources (PGR) are rapidly being molecularly characterized using various marker systems (Larsen et al., 2018; Mascher et al., 2019; Milner et al., 2019). In order to effectively manage and use plant genetic resources, different methods could be employed to select a core collection (Jeong et al., 2017). Harnessing marker information to select core collections based on aspects of genetic diversity such as pairwise dissimilarity, allelic richness, or heterozygosity is feasible today (van Heerwaarden et al., 2013). Core collections use distance metrics to quantify the similarity of two accessions, based on genetic marker data or phenotypic traits (De Beukelaer et al., 2018). Different distance metrics or traits can be applied to generate core sets that are specific for a particular purpose e.g. maximizing the genetic diversity in a trait of interest. The Core Hunter software is a core set selection tool known for its flexibility to sample diverse, representative subsets from large germplasm collections with minimal redundancy (http://www.corehunter.org). Three different main versions of Core Hunter have been released. Core Hunter 3 was introduced by De Beukelaer et al. (2018) as a multi-purpose tool for selecting core subsets. For this purpose, Core Hunter 3 uses local search algorithms to provide subsets based on several distance metrics and allelic abundance. The software is capable of combining distances, entry-to-nearest-entry (E-NE) and accession-to-nearest-entry (A-NE) computations (De Beukelaer et al., 2018). Based on genetic markers, genetic differences between genotypes are calculated to evaluate the core subsets. Different methods for calculating distances are implemented. The user can either provide a genetic distance matrix which is estimated using a suitable measure such as Modified Roger's distance (Wright, 1984). On the other hand, the user can provide phenotypic traits, which are then evaluated with Gower's Distance to derive a phenotypic distance matrix (Gower, 1971).

However, for the selection of core collections, there are general clustering methods, i.e. hierarchical and partial clustering using different subtypes and algorithms to identify clusters (Kaski, 1997). Here, the focus is on partial clustering. Partial clustering comprises two clustering approaches: k-means (MacQueen, 1967) and *k*-medoids (Kaufman and Rousseeuw, 1987). *K*-medoids is known as a modified version of k-means. Both methods minimize the distance between data points within a cluster to the respective cluster center (Block et al., 2019). The

main difference between methods lies in the handling of the cluster centers: While in *k*-medoids the cluster center needs to be a real object of the collection, the cluster center is an average of all cluster members in k-means and does not need to be a real object of the collection. To distinguish the two types of cluster centers, they are either called medoids (*k*-medoids) or centroids (*k*-means). Usually, *k*-medoids is considered the more robust algorithm in terms of clustering, as it is less sensitive to outliers compared to *k*-means (Park et al., 2006; Park and Jun, 2009). *K*-medoids has been used in various applications: in genetics (Broin et al., 2015), in geography (Bernábe-Loranca et al., 2014), in analyses to predict the popularity of television programs (Zhu et al., 2017), and as a decision support system in the fashion industry (Monte et al., 2013). The availability of genotypic information for different genotypes allows clustering the genotypes based on similarity or dissimilarity.

High-throughput technologies, such as next generation sequencing (NGS) or array-based technologies, offer the possibility of generating comprehensive genotype data for entire plant genomes in a short time and with high accuracy (Varshney et al., 2009). Such genotype information is also frequently used to identify marker–trait association in quantitative trait locus (QTL) mapping and genome wide association studies (GWAS) (Wang et al., 2014). The development of single nucleotide polymorphism (SNP) data has significantly increased the knowledge of genome diversity. On the other hand, advances in NGS reduced the cost of DNA sequencing, which made genotyping-by-sequencing (GBS) possible for species with high diversity and large genomes (He et al., 2014).

Several genotyping array based platforms for wheat have been published (Ganal et al., 2019). First, Cavanagh et al. (2013) developed a 9K Illumina iSelect SNP array with 9,000 SNPs. In 2014, Wang et al. (2014) reported a 90K Illumina iSelect SNP array based on the 9K array technology. The third array based platform for wheat genotyping was the Affymetrix Axiom 820K SNP array presented by Winfield et al. (2016). With this array it was possible to genotype not only hexaploid wheat but to detect and track introgressions from different sources. A subset of the markers used on this 820K array were then used to develop the Axiom 35K SNP array (Allen et al., 2017), which was specifically targeted at the elite wheat germplasm. Here we present the 15K array, a new and optimized platform containing a set of 12,908 optimized SNP markers mainly originating from the 90K chip design. This subset offers a cost-effective alternative to the 90K array.

In this paper two different methods, namely *k*-medoids and Core Hunter 3, were applied to select different sizes of core collections from a large set of wheat genetic resources and were compared to identify the most appropriate method.

MATERIAL AND METHODS

Development of the 15K Wheat Infinium Array

The 15K wheat Infinium array has been developed mainly based on genotyping data for more than 2,000 wheat genotypes consisting of European and world-wide lines, that have been generated with the 90K wheat Infinium array (Wang et al., 2014) at TraitGenetics. The selection steps that were applied to create the 15K array are as follows:

1. Based on the raw genotyping data, all markers were surveyed for marker quality during the cluster file development using the Illumina GenomeStudio software (Illumina, San Diego, USA). Markers with clearly differentiated clusters were identified independently whether the markers were genome-specific (Ganal et al., 2012).
2. Genetic mapping data (Wang et al., 2014) were used together with additional mapping data generated from the ITMI DH population (Sorrells et al., 2011) for selecting markers that are evenly distributed throughout the genetic map of the three (A, B, D) wheat genomes.
3. Using the marker order determined by the genetic mapping, additional markers were integrated in case they were in perfect linkage disequilibrium with at least one other mapped marker.
4. Haplotype blocks were defined as containing markers in perfect linkage disequilibrium over all investigated wheat lines. From each larger haplotype block especially in the centromeric regions, one or two markers were selected based on the marker quality defined by Wang et al. (2014).
5. The markers from the 90K array were supplemented by 383 additional markers from an unpublished 12K wheat Infinium array previously developed by TraitGenetics for haplotype blocks that were not identified using the 90K markers.
6. Finally, a set of 27 public markers derived from candidate genes for major wheat phenology traits has been added.

In total, 15,000 markers were submitted for array design to Illumina of which 12,908 markers remained after array manufacturing and an additional genotyping round of 384 wheat lines to identify low quality markers. These were used for the development of a cluster file for allele calling. These functional markers are listed in **Supplementary Table S1** which also includes information about the origin (90K or 12K or candidate gene) and the respective context sequence.

Plant Material

In this study, a collection of 890 winter wheat genotypes was used for the development of a small genetically diverse core collection. The 890 genotypes were collected from five different collections, which had been used in different studies at the Julius Kuehn Institute, Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance (JKI-RS). Ninety two were evaluated under drought stress and well-watered conditions in the presence and absence of mycorrhizae to identify QTLs involved in response to mycorrhizae under drought stress condition (collection 1) (Lehnert et al., 2018). Babben et al. (2018) and Soleimani et al. (in preparation) evaluated 284 genotypes to identify genome regions associated with frost tolerance (collection 2). A set of 40 genotypes was tested for resistance against soil borne viruses (collection 3). These three collections were genotyped by using the 90K

Illumina iSelect array (Wang et al., 2014), with the exception of five genotypes from collection 3, which were genotyped using the 15K Infinium array. Furthermore, 220 genotypes were evaluated under two different nitrogen concentrations [collection 4, (Voss-Fels et al., 2019)], and 254 genotypes were inoculated with wheat dwarf virus to select genotypes tolerant against this virus (collection 5), respectively. These genotypes were genotyped by using the 15K Infinium array.

As two different platforms (15K and 90K) were used for genotyping the wheat genotypes, only common markers (markers which were detected by the 15K and 90K array approach) were used for further analyses. A principal coordinate analysis (PCoA) was conducted with the package 'ape' (Paradis and Schliep, 2018) in the R statistical environment based on the Modified Roger's distance (MRD) matrix to visualize the genetic diversity in the five collections.

Placement of SNP Array Marker Sequences Onto the Pseudomolecule Reference Sequence

The published reference genome of the bread wheat cultivar Chinese Spring (the IWGSC RefSeq) and the genome annotation were downloaded (Appels et al., 2018). SNP array marker sequences were split at the polymorphic site with a custom awk script and turned into paired-end style sequencing reads, effectively reverse complementing one of the reads. These artificial paired-end reads were then mapped to the bread wheat pseudomolecule reference sequence with BWA mem (version 0.7.13) with -M parameter for highlighting of secondary alignments (Li and Durbin, 2009; Li, 2013). Alignments were converted to BAM format with SAMtools (version 1.6) (Li et al., 2009). Unmapped reads and secondary alignments were discarded and remaining high quality alignments (MAPQ ≥ 20) were transformed to BED format with BEDtools (version 2.8) keeping the CIGAR string (Quinlan and Hall, 2010; Quinlan, 2014). Filtered alignments were then checked for consistency with a custom Java program. Briefly, reads without a mapped mate, pairs of reads that do not map exactly one nucleotide apart, and mapped reads where the SNP position was an unknown nucleotide ('N') were removed. Afterwards, all mapped markers were evaluated on the 890 genotypes. Markers with equal or more than 30% of missing data as well as monomorphic markers were removed from further analysis. Duplicate markers and markers mapping to the same physical position were removed as well and only the initial marker was kept. The filtered marker data were used for SNP imputation by applying the software package Beagle version 4.1 (Browning and Browning, 2007; Browning and Browning, 2009). Imputed marker data were filtered for minor allele frequency (MAF) ≥ 5%, and heterozygosity ≤ 12.5%, resulting in a set of 7,672 SNP markers used for subsequent analyses.

K-Medoids Clustering

Based on the Modified Roger's distance (MRD) matrix, 890 genotypes were clustered into 178 and 320 groups by using the *k*-medoids clustering method (Kaufman and Rousseeuw, 1987). *K*-medoids clustering was conducted by using the cluster package (version 2.1.0) and PAM method in the R statistical environment (Maechler et al., 2012; RDevelopment CORE TEAM, 2015).

Core Hunter 3

Two different genetic distances, 1) MRD (Roger, 1972; Wright, 1984), 2) and Cavalli-Sforza and Edwards (CSE) distance (Cavalli-Sforza and Edwards, 1967) were applied to calculate different core sets. In total, 14,000 different core sets were determined (two sizes times seven different settings times 1000 iterations in Core Hunter 3). Different approaches for calculating core sets in Core Hunter 3 were used, *i.e.*:

a. Average Entry-to-Nearest-Entry distance (E-NE) (Odong et al., 2013): This is the mean distance between all selected accessions and their closest other selected accession. Maximizing this measure yields high diversity in the core collection expressed through maximum dissimilarity of selected core accessions (De Beukelaer et al., 2018). Both genetic distances (MRD and CSE) were applied for calculating these core sets.
b. Average Accession-to-Nearest-Entry distance (A-NE) (Odong et al., 2013): The A-NE considers the mean distance between each accession in the whole collection and the closest selected accession. Minimizing this measure yields core collections that maximally represent all individual accessions from the full collection (De Beukelaer et al., 2018). Both genetic distances (MRD and CSE) were applied for calculating these core sets.
c. Shannon's diversity index (Shannon, 1948): Shannon's diversity index is an appropriate measure when forming core subsets that attempt to retain as many rare alleles as possible, regardless of their co-location within loci (Thachuk et al., 2009). The Shannon diversity index achieves its highest value when each allele exists only once in the whole data set being measured.
d. Expected heterozygosity (Berg and Hamrick, 1997): The expected proportion of heterozygous loci on the other hand, specifically considers diversity within each locus. Intuitively, since each locus contributes equally to the overall value of this measure, core subsets selected using this measure are less likely to be homozygous for a number of different loci than core subsets selected with Shannon's Diversity index (Thachuk et al., 2009).
e. Allele coverage: The percentage of marker alleles observed in the full collection that are retained in the core. This is a simple measurement, which indicates the percentage of retained alleles in the core set relative to the whole population. This method is particularly useful for selecting core sets to preserve alleles in gene and seed banks (Thachuk et al., 2009).

RESULTS

The overlap between the 15K and 90K arrays resulted in 8221 SNP markers that could be mapped to unique positions in the

reference wheat genome sequence. Of these markers, the majority (45%) mapped to sub-genome B, followed by sub-genome A with 39%, while the lowest proportion (15%) was located on sub-genome D. Less than 1% of markers were mapped to sequences located to chromosome 'unknown', an artificial chromosome consisting of sequences that could not be assigned to any chromosome yet. Among the chromosomes, the highest and lowest number of mapped markers was identified on chromosomes 5B and 4D with 595 and 62 markers respectively. The number of mapped markers per chromosome is listed in **Table 1**. To understand the effects on observed versus expected heterozygosity based on the array system, a set of 48 wheat accessions was analyzed by genotyping them with the 15K and 90K array. During this comparison no significant differences between array systems was detected (**Figure S1**).

The quality check of the markers resulted in a set of 7,672 polymorphic, informative markers (**Figure 1**). These markers were placed at unique positions on the reference genome sequence of bread wheat (*cv.* Chinese Spring). This final set of markers was used for further analyses.

Furthermore, a principal coordinate analysis (PCoA) was performed (**Figure S2**). The first and the second principle coordinates (PCs) explained 9.5 and 4.2% of the total variance and were used to graphically display the results. The analysis showed that most genotypes from the different collections were not clearly separated from each other. Although clusters of genotypes from collections can be observed, outliers from each collection can also be found near or within clusters of other collections. Most genotypes belong to the collections 2 and 5.

Comparing Different Core Sets

In total, 178 and 320 genotypes were selected by *k*-medoids clustering and Core Hunter 3, respectively. Core Hunter 3 uses random seeds and a non-deterministic algorithm to arrive at a solution after a time (or alternatively step) threshold has been reached. Similarly, the *k*-medoids algorithm as implemented in the PAM function inside the R library 'cluster' also works non-deterministic. However, in the so-called build phase the program chooses a good initial set of medoids. In our tests given our population and MRD matrix, it always produced the same core collection. Therefore, we randomly sampled initial medoids and gave these to the PAM function as input parameters allowing to compare the stability of the obtained results with those from Core Hunter 3.

TABLE 1 | Distribution of uniquely mapped markers on the reference genome sequence from the 15K SNP array.

Chromosome	Wheat genome			Total
	A	B	D	
1	451	580	270	1,301
2	480	710	289	1,479
3	415	556	142	1,113
4	287	258	62	607
5	508	595	186	1,289
6	485	530	169	1,184
7	546	485	144	1,175
Total	3,172	3,714	1,262	8,148
Unknown		73		8,221

Our goal was to assess the results obtained through a large number of iterations (n = 1000) to get information on 1) the stability of the methods, 2) the influence of the size of the core collection size, and 3) which method performs best for the two main objectives to form core collections: CC-I and CC-X.

For testing the stability of the different methods implemented in Core Hunter 3 and *k*-medoids we performed an empirical cumulative distribution analysis with the function 'ecdf' in the R statistical environment. We evaluated the resulting core sets by looking at the composition of entries in 1,000 runs per method and two different core set sizes (178 and 320). For the goal of observing the gain from using any core selection program, we also constructed 1000 random sets per core set size using the R function 'sample'. The stability results for all tested methods demonstrated similar behavior for both core set sizes of 178 and 320 genotypes (**Figure 2**). Taking into account the definition of stability (Higham, 2002; Atkinson and Han, 2005; Soleimani and Weiner, 2018), a method returns stable results if all genotypes are either never or always selected. In contrast a method returns unstable results if all genotypes are uniformly selected. However, stability is not a binary feature, it is much more continuous. The stability test was characterized by the frequency of a genotype selected by a method as an entry into a core collection. The ecdf of a stable method should be close to the grey horizontal dotted line, while the ecdf of an unstable method should be close to the grey vertical dashed line. Based on the observed results, Shannon's diversity and expected heterozygosity in all 1,000 runs showed a high number of entries in the core sets that were common between runs and can therefore be considered stable methods (**Figure 2**). On the basis of the stability analysis we obtain a ranking of the applied methods according to increasing stability: A-NE, E-NE, *k*-medoids, Shannon's diversity and expected heterozygosity. Both the random and allele coverage sets, on the other hand, showed a very unstable behavior.

To evaluate the quality of selected core sets, we calculated two average distances as proposed by Odong et al. (2013). The average A-NE result varied between 0.25 to 0.39 and 0.17 to 0.29 for a core set size of 178 and 320 genotypes, respectively (**Figures 3A, B**). The lowest average A-NE was observed for *k*-medoids, and also the average Accession-to-Nearest-Entry method (A-NE) showed low values for average A-NE. Both Shannon's diversity (SD) and expected heterozygosity (EH) showed high values for average A-NE and therefore performed worse compared to the other methods. Based on results obtained for average A-NE, the methods *k*-medoids and A-NE were best suited to represent the original population due to the smallest value for average A-NE (**Figures 3A, B**) for both sizes of core sets.

Furthermore, our results for average E-NE calculation for both sizes of core sets showed that the method based on Entry-to-Nearest-Entry distance (E-NE) performed better to represent extreme genotypes compared to other core sets, as the obtained average E-NE showed the highest value among all analyzed core sets (**Figures 4A, B**). The methods based on Shannon's diversity (SD) and expected heterozygosity (EH) showed the lowest values for the average E-NE. Therefore, based on observed results, two

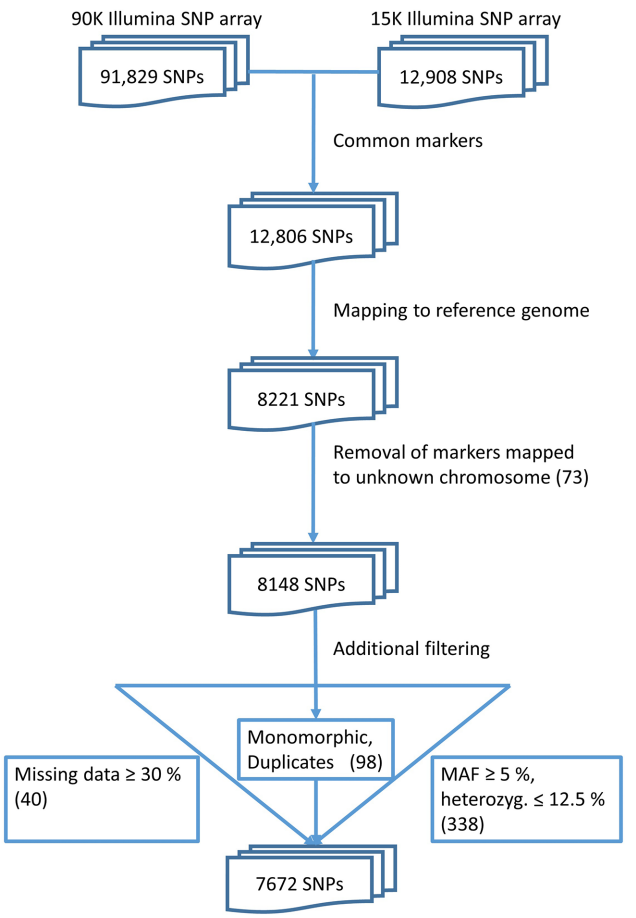

FIGURE 1 | Data pre-processing finally yields a set of 7,672 markers.

core set methods (SD and EH) indicate an insufficient representation of the extreme genotypes from the original population in both sizes of core sets.

The two genetic distance metrics, MRD and CSE, that were used for the two core selection methods A-NE and E-NE produced very similar results throughout the different evaluations (**Figure 5**) and for the sake of simplicity only the results obtained by using MRD are shown in **Figures 2**–**4**.

DISCUSSION

The development and use of molecular markers has expanded our knowledge to better understand cereal genetics. High-throughput SNP array genotyping allows genotyping thousands of markers in parallel. This technique has been applied in recent years for small grain cereals such as barley, wheat, rye, and oats (Ganal et al., 2019). The 90K Illumina Infinium array is currently the most widely used genotyping array in wheat. However, this genotyping array is quite expensive on a price per sample base and creates a large set of redundant data (Ganal et al., 2019). Subsequently, the Affymetrix Axiom 820K SNP array was developed to genotype wheat and to detect and track introgressions. Later, this technology was used for the development of the Axiom 35K SNP array. In this study, we also used the new 15K Illumina Infinium array with 12,908 functional markers that contains mainly high quality and informative markers. The overlap between the two array platforms (15K and 90K) is 12,806 markers. The 15K genotyping array with a lower number of markers is a cost-effective option for genotyping experiments that still provides high resolution data.

Breeders seek to improve yield performance by exploiting favorable traits associated with tolerance against biotic and abiotic stress (Pandey et al., 2017). Germplasm collections from major crops have increased in size and number worldwide (Brown et al., 1997). Genebanks play an important role in securing genetic diversity for future use. They are distributed around the world and preserve the genetic diversity in crop species (Shands, 1990; Fowler and Hodgkin, 2004).

The increase in the size of germplasm collections leads to problems and complications in the characterization, evaluation, utilization and maintenance of germplasm. The first approach to reduce the size of large collections and to select core sets of these collections was defined by Frankel (1984). Core collections became important due to the demand for more efficiency in the characterization and utilization of collections stored in genebanks (Odong et al., 2013). Different methods are available to create core collections for varying purposes with respect to phenotypic and genotypic data. These methods could be used to select genetically diverse genotypes for carrying out different scientific research before a large number of genotypes are phenotyped, thus excluding genotypes that would show the same behavior. Therefore, by eliminating the need for an additional phenotyping step, these approaches could accelerate research experiments and breeding programs. Molecular markers are widely used to unlock the genetic diversity of germplasm collections. Odong et al. (2013) pointed out the role of genetic differentiation in marker data, which has a significant impact on core selection methods.

Different algorithms are known for the generation of core sets, and comparisons between different algorithms have been made in previous studies. For example, Thachuk et al. (2009) compared three different algorithms (D-method, MSTRAT and PowerCore) with Core Hunter to select core sets in a maize population. The comparisons confirmed that Core Hunter performed better than other methods in creating core sets with higher genetic diversity. Also, Core Hunter was able to select significantly smaller core subgroups that retained all unique alleles from an original collection than the other algorithms. In our study, we used the same genetic distance and genetic diversity indices as Thachuk et al. (2009) to compare *k*-medoids and Core Hunter 3 for core collection selection.

In the present study, we conducted a stability test for six methods comprising allele coverage (AC), expected heterozygosity (EH), Shannon's diversity (SD), A-NE, E-NE and *k*-medoids to analyze their reproducibility. Based on the definition of Higham (2002) and Atkinson and Han (2005), SD and EH, were more stable than other methods. A-NE and E-NE

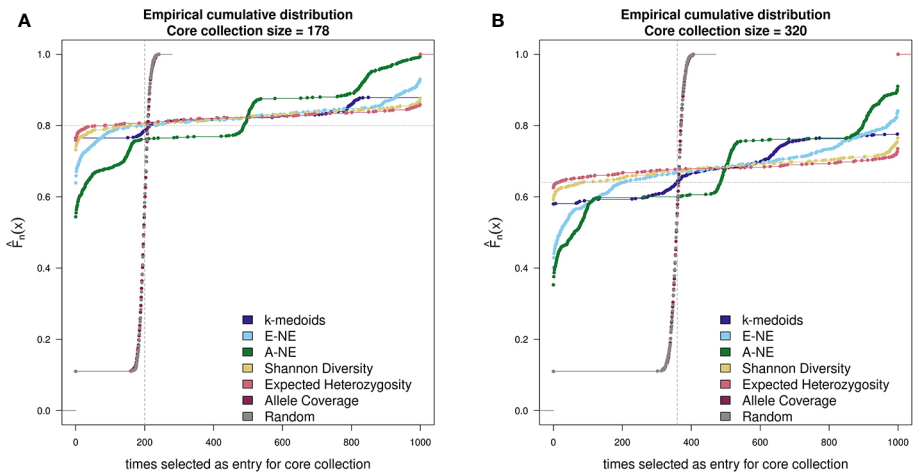

FIGURE 2 | Comparison of stability test with 1,000 runs between *k*-medoids derived core set, seven core sets derived by Core Hunter 3 and randomly selected core sets. **(A)** depicts the stability results for the core set containing 178 genotypes, while **(B)** depicts the stability results for the core set containing 320 genotypes. Methods with a low gradient are considered to be stable; large gradients, on the other hand, show a high degree of variability. Two gray helper lines have been added for easier visual interpretation of results. The dotted horizontal line indicates stable results, while the dashed vertical lines shows instability.

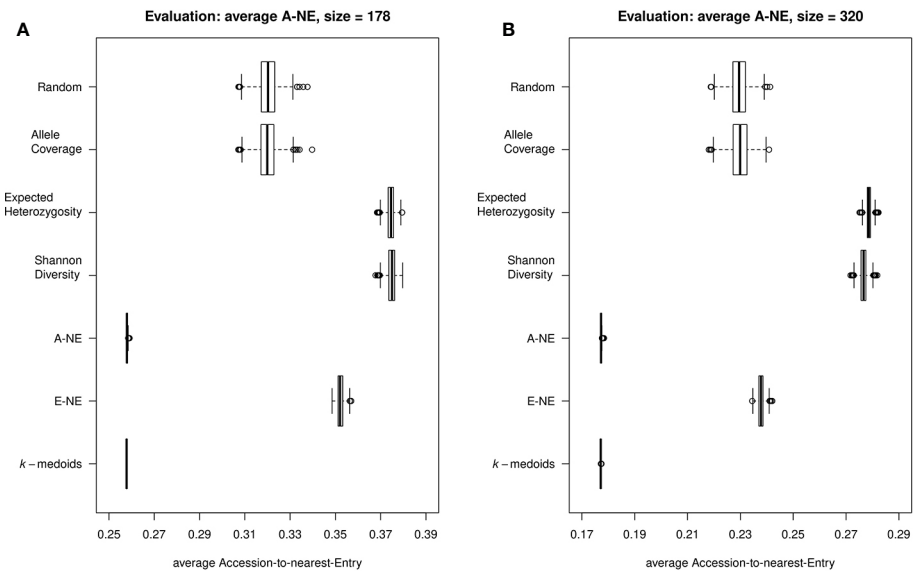

FIGURE 3 | Quality of core collections. Displayed are the average distances between each of the 890 accessions to the nearest entry of the respective core set (A-NE) for core collections of different sizes. **(A)** shows core sets of size 178, while **(B)** shows core sets of size 320. A low average distance is favorable to obtain a good representation of the original collection.

methods provided by Core Hunter 3 as well as *k*-medoids can be classified as stable methods for the selection of core collections. AC showed a highly unstable performance when selecting core sets and should be avoided when core sets should be reproducible (as it also highly resembled the random selected sets).

In the present study, two genetic metrics were applied to assess the quality of different core set selection methods (Odong et al., 2013). For the evaluation of CC-I core sets, the calculation of the average A-NE is a suitable method. For such an objective the average A-NE value should be as small as possible. An average A-NE value equal to zero indicates a minimal distance between genotypes and thus the maximum representation of the genotype in the core collection. Based on this definition, the *k*-medoids and A-NE derived core sets did the best job to achieve maximum genetic diversity of genotypes with the lowest average value of A-NE observed. On the other hand, a good criterion for the evaluation of CC-X core sets is to maximize the average E-NE. The E-NE method describes how genetically diverse the

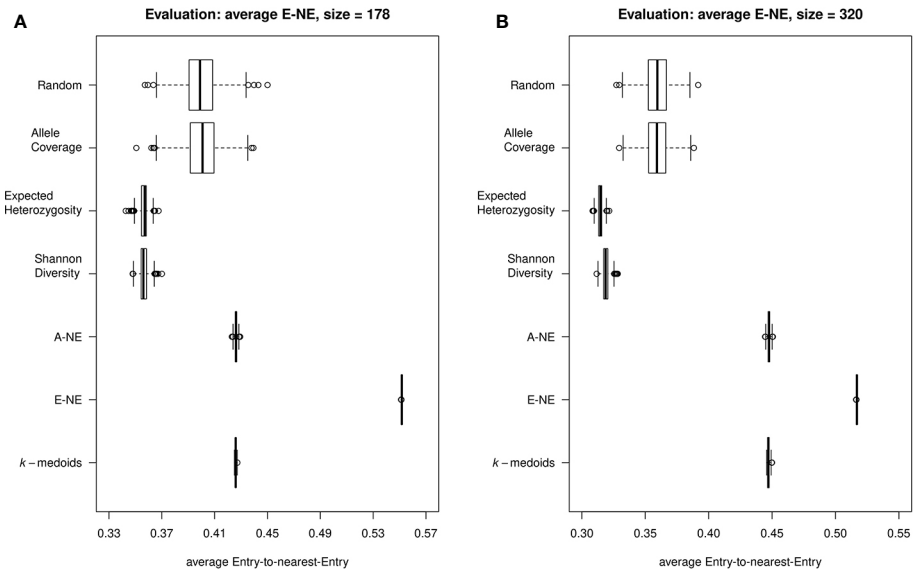

FIGURE 4 | Quality of the core sets. Displayed are the average distances between each of the entries to the nearest entry of the respective core set (E-NE) for core sets of different sizes. **(A)** shows core sets of size 178, while **(B)** shows core sets of size 320. A high average distance is favorable to obtain a good representation of the extreme genotypes of the original collection.

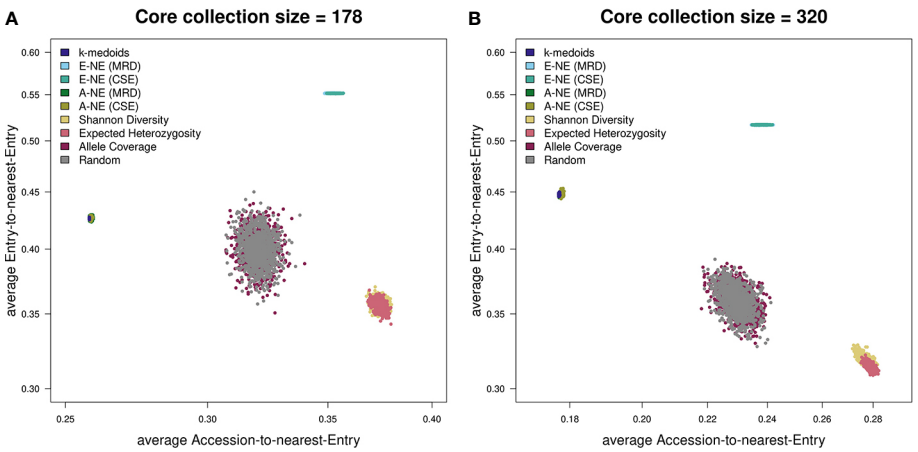

FIGURE 5 | Scatterplots showing both average A-NE and average E-NE for the observed core collections for different sizes. **(A)** shows core collections of size 178, while **(B)** shows core collections of size 320. As already indicated by the stability test (**Figure 2**), the core collections from the type allele coverage show a large variance in their distribution. The Shannon diversity and expected heterozygosity methods seem to produce core collections of similar quality. The same seems to be true for *k*-medoids and A-NE methods. A theoretically optimal core collection would be located in the upper left corner of the plot.

entries into the core set are to each other. Therefore, the best possible core set for CC-X strategy has the highest average E-NE. In our tests, the average Entry-to-Nearest-Entry (E-NE) core collections compared to other core set methods performed best in this category. However, it is not surprising that A-NE derived core collections yield good results for CC-I and E-NE derived core collections yield good results for CC-X.

For a final assessment of core selection methods, we evaluated and combined the results of the stability test and the quality of core selection on the basis of average A-NE/E-NE. Based on the stability tests, the most stable core selection methods are Shannon's diversity (SD) and expected heterozygosity (EH). While these two core selection methods showed less good results for the average A-NE and the average E-NE for different purposes (CC-I and CC-X) of core collections, they should therefore not be considered superior to the other core selection methods. Although *k*-medoids is a general clustering method and is not specifically designed for creating core

collections, it proved to be one of the better methods for creating CC-I core sets due to its small average A-NE value. Based on our results from the evaluation with average E-NE, *k*-medoids also proved to be an adequate method for the generation of CC-X core sets. Interestingly, the A-NE based core selection methods showed very similar profiles to the *k*-medoids method in both average A-NE and average E-NE evaluation, but were somewhat more unstable in the stability test (**Figures 3**–**5**).

CONCLUSION

In the present study, we used the wheat 90K Infinium array together with an optimized 15K Infinium array with 12,908 informative markers. Compared to the 90K array, the 15K array is a cost-effective platform for research and plant breeding programs that generates high quality data. We selected core collections of 178 and 320 genotypes from a collection of 890 wheat genotypes using *k*-medoids and Core Hunter 3. Two genetic distances and three indices of genetic diversity were used to establish core collections and the results were compared to determine the best approach for a large population of diverse genotypes. Our results support the conclusion that choosing either MRD or CSE as genetic distance has little to no observable effect on the selection of core collections using A-NE and E-NE in Core Hunter 3. In addition, *k*-medoids and Accession-to-Nearest-Entry (A-NE) are appropriate methods to select a uniform representation of the original population (CC-I). However, if the purpose of generating a core collection is to construct a core set based on the extremes of the relevant traits (CC-X), the method Entry-to-Nearest-Entry (E-NE) showed the best results. Furthermore, both *k*-medoids and A-NE methods seem to be a good compromise when trying to combine the goals of CC-I and CC-X (**Figure 5**). Finally, A-NE, E-NE and *k*-medoids yield stable results if started multiple times independently.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in Zenodo, DOI: 10.5281/zenodo.3905912.

AUTHOR CONTRIBUTIONS

FO and DP conceived and designed the experiments, collected all genotypic data from five different collections for 890 genotypes of wheat. HL, JK, SN, SB, and BS performed the statistical analyses on the data. JP and MG provided the 15K array design and data. SB and BS wrote the initial draft. SB, BS, DP, HL, JK, and SN interpreted the data. All authors contributed to the article and approved the submitted version.

FUNDING

This research was financially supported by a grant (project 031B0186D) from the German Federal Ministry of Education and Research, Bundesministerium für Bildung und Forschung (BMBF).

ACKNOWLEDGMENTS

We would like to thank PD Dr. Andreas Börner, Dr. Holger Zetzsche and Dr. Antje Habekuß for providing 15K and 90K iSelect array data and to Thomas Berner for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01040/full#supplementary-material

FIGURE S1 | Comparison of the effect of the choice of the array system on the observed and expected heterozygosity. A set of 48 wheat accessions was genotyped on both 15K and 90K Illumina Infinium arrays and both observed and expected heterozygosity were calculated. Panel **(A)** shows the observed heterozygosity (Ho) on 15K and 90K arrays, panel **(B)** shows the expected heterozygosity (He) on 15K and 90K arrays. The relative frequency of observed heterozygosity (Ho) is shown in panels **(C)** (90K array) and **(D)** (15K array), while the relative frequency of expected heterozygosity (He) is shown in panels **(E)** (90K array) and **(F)** (15K array).

FIGURE S2 | Principal coordinate analysis (PCoA) indicating genetic diversity over five different collections for a total population of 890 wheat genotypes.

REFERENCES

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., et al. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (Triticum aestivum). *Plant Biotechnol. J.* 15, 390–401. doi: 10.1111/pbi.12635

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, 661–66+. doi: 10.1126/science.aar7191

Atkinson, K., and Han, W. (2005). *Theoretical numerical analysis* (New York: Springer).

Babben, S., Schliephake, E., Janitza, P., Berner, T., Keilwagen, J., Koch, M., et al. (2018). Association genetics studies on frost tolerance in wheat (Triticum aestivum L.) reveal new highly conserved amino acid substitutions in CBF-A3, CBF-A15, VRN3 and PPD1 genes. *BMC Genomics* 19, 409. doi: 10.1186/s12864-018-4795-6

Berg, E. E., and Hamrick, J. (1997). Quantification of genetic diversity at allozyme loci. *Can. J. For. Res.* 27, 415–424. doi: 10.1139/x96-195

Bernábe-Loranca, B., Gonzalez-Velázquez, R., Olivares-Benítez, E., Ruiz-Vanoye, J., and Martínez-Flores, J. (2014). Extensions to K-Medoids with Balance Restrictions over the Cardinality of the Partitions. *J. Appl. Res. Technol.* 12, 396–408. doi: 10.1016/S1665-6423(14)71621-9

Block, K., Trumm, S., Sahitaj, P., Ollinger, S., and Bergmann, R. (2019). *Clustering of Argument Graphs Using Semantic Similarity Measures. Presented at the Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)* (Cham: Springer), 101–114.

Broin, P.Ó., Smith, T. J., and Golden, A. A. (2015). Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach. *BMC Bioinf.* 16, 22. doi: 10.1186/s12859-015-0450-2

Brown, A., Brubaker, C., and Grace, J. (1997). Regeneration of germplasm samples: wild versus cultivated plant species. *Crop Sci.* 37, 7–13. doi: 10.2135/cropsci1997.0011183X003700010002x

Brown, A. (1989). Core collections: a practical approach to genetic resources management. *Genome* 31, 818–824. doi: 10.1139/g89-144

Browning, B. L., and Browning, S. R. (2007). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc* 31, 365–375. doi: 10.1002/gepi.20216

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005

Cavalli-Sforza, L. L., and Edwards, A. W. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* 21, 550–570. doi: 10.1111/j.1558-5646.1967.tb03411.x

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.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

De Beukelaer, H., Davenport, G. F., and Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinf.* 19, 203. doi: 10.1186/s12859-018-2209-z

Fowler, C., and Hodgkin, T. (2004). Plant genetic resources for food and agriculture: assessing global availability. *Annu. Rev. Env. Resour.* 29, 143–179. doi: 10.1146/annurev.energy.29.062403.102203

Frankel, O. (1984). "Genetic perspectives of germplasm conservation" in *Genetic Manipulation: Impact on Man and Society.* Eds. WK Arber, K Llinensee, WJ Peacock, et al. (Cambridge: Cambridge University Press), 61, 161–170.

Ganal, M. W., Polley, A., Graner, E.-M., Plieske, J., Wieseke, R., Luerssen, H., et al. (2012). Large SNP arrays for genotyping in crop plants. *J. Biosci.* 37, 821–828. doi: 10.1007/s12038-012-9225-3

Ganal, M. W., Plieske, J., Hohmeyer, A., Polley, A., and Röder, M. S. (2019). "High-Throughput Genotyping for Cereal Research and Breeding," in *Applications of Genetic and Genomic Research in Cereals* (Woodhead Publishing), 3–17.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–871. doi: 10.2307/2528823

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5, 484. doi: 10.3389/fpls.2014.00484

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*, (University City, Philadelphia: Society for Industrial and Applied Mathematics).

Jeong, S., Kim, J.-Y., Jeong, S.-C., Kang, S.-T., Moon, J.-K., and Kim, N. (2017). GenoCore: A simple and fast algorithm for core subset selection from large genotype datasets. *PloS One* 12, e0181420. doi: 10.1371/journal.pone.0181420

Kaski, S. (1997). *Data exploration using self-organizing maps. Presented at the Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82*, (Espoo).

Kaufman, L., and Rousseeuw, P. J. (1987). "Clustering by Means of Medoids" in *Statistical Data Analysis Based on the L1–Norm and Related Methods.* (Y. Dodge, Dü.) Reports of the Faculty of Mathematics and Informatics. Delft University of Technology, 405, 405–416.

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S., et al. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PloS One* 13, e0201889. doi: 10.1371/journal.pone.0201889

Lehnert, H., Serfling, A., Friedt, W., and Ordon, F. (2018). Genome wide association studies reveal genomic regions associated with the response of wheat (Triticum aestivum L.) to mycorrhizae under drought stress conditions. *Front. Plant Sci.* 9, 1728. doi: 10.3389/fpls.2018.01728

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

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

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997, 1–3.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," *Presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Oakland, CA, USA), 281–297.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2012). *Cluster: cluster analysis basics and extensions. R Package Version 1.14.2,* Available at: http://CRAN.R-project.org/package=cluster.

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* 1, 1076–1081. doi: 10.1038/s41588-019-0443-6

Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., et al. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51, 319–326. doi: 10.1038/s41588-018-0266-x

Monte, A., Soares, C., Brito, P., and Byvoet, M. (2013). "Clustering for Decision Support in the Fashion Industry: A Case Study," in *Advances in Sustainable and Competitive Manufacturing Systems* (Heidelberg: Springer), 997–1008.

Odong, T., Jansen, J., Van Eeuwijk, F., and van Hintum, T. J. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theor. Appl. Genet.* 126, 289–305. doi: 10.1007/s00122-012-1971-y

Pandey, P., Irulappan, V., Bagavathiannan, M. V., and Senthil-Kumar, M. (2017). Impact of combined abiotic and biotic stresses on plant growth and avenues for crop improvement by exploiting physio-morphological traits. *Front. Plant Sci.* 8, 537. doi: 10.3389/fpls.2017.00537

Paradis, E., and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Park, H.-S., and Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* 36, 3336–3341. doi: 10.1016/j.eswa.2008.01.039

Park, H.-S., Lee, J.-S., and Jun, C.-H. (2006). A K-means-like algorithm for K-medoids clustering and its performance. *Proceedings of The 36th International Conference on Computers and Industrial Engineering*. 2006 June 20–23; 1222–1231.

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

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinform.* 11.12, 1–11.12. 34. doi: 10.1002/0471250953.bi1112s47

RDevelopment CORE TEAM, R (2015). *R: A language and environment for statistical computing*, Vienna, Austria. Available at: https://www.R-project.org/

Roger, J. (1972). *Measure of genetic similarity and genetic distance. Studies in genetics VII* Vol. 7213 (Austin, Texas: University of Texas Publication), 145–153.

Shands, H. L. (1990). Plant genetic resources conservation: the role of the gene bank in delivering useful genetic materials to the research scientist. *J. Hered.* 81, 7–10. doi: 10.1093/oxfordjournals.jhered.a110928

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Soleimani, B., and Weiner, R. (2018). Superconvergent IMEX peer methods. *Appl. Numer. Math.* 130, 70–85. doi: 10.1016/j.apnum.2018.03.014

Sorrells, M. E., Gustafson, J. P., Somers, D., Chao, S., Benscher, D., Guedira-Brown, G., et al. (2011). Reconstruction of the Synthetic W7984× Opata M85 wheat reference population. *Genome* 54, 875–882. doi: 10.1139/G11-054

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., and Davenport, G. F. (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinf.* 10, 243. doi: 10.1186/1471-2105-10-243

van Heerwaarden, J., Odong, T., and van Eeuwijk, F. (2013). Maximizing genetic differentiation in core collections by PCA-based clustering of molecular marker data. *Theor. Appl. Genet.* 126, 763–772. doi: 10.1007/s00122-012-2016-2

Varshney, R. K., Nayak, S. N., May, G. D., and Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530. doi: 10.1016/j.tibtech.2009.05.006

Voss-Fels, K. P., Stahl, A., Wittkop, B., Lichthardt, C., Nagler, S., Rose, T., et al. (2019). Breeding improves wheat productivity under contrasting agrochemical input levels. *Nat. Plants* 5, 706–714. doi: 10.1038/s41477-019-0445-5

Wang, S. C., Wong, D. B., Forrest, K., Allen, A., Chao, S. M., 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

Wang, X., Bao, K., Reddy, U. K., Bai, Y., Hammar, S. A., Jiao, C., et al. (2018). The USDA cucumber (Cucumis sativus L.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development. *Hortic. Res.* 5, 64. doi: 10.1038/s41438-018-0080-8

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

Wright, S. (1984). *Evolution and the genetics of populations, volume 3: experimental results and evolutionary deductions* (Chicago: University of Chicago press).

Zhu, C., Cheng, G., and Wang, K. (2017). Big data analytics for program popularity prediction in broadcast TV industries. *IEEE Access* 5, 24593–24601. doi: 10.1109/ACCESS.2017.2767104

Conflict of Interest: The authors JP and MG have competing commercial interests as members of TraitGenetics GmbH which is a company that offers marker development and analysis (including this array) for commercial purposes. This does not alter the authors' adherence to sharing all data and materials. There are no further products in development or marketed products or patents to declare.

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.

frontiers in Plant Science

METHODS
published: 30 September 2020
doi: 10.3389/fpls.2020.511768

R/UAStools::plotshpcreate: Create Multi-Polygon Shapefiles for Extraction of Research Plot Scale Agriculture Remote Sensing Data

Steven L. Anderson II[†] *and Seth C. Murray*[*]

Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Roger Deal,
Emory University, United States

Reviewed by:
Xu Wang,
Kansas State University, United States
David Shaner LeBauer,
University of Arizona, United States

****Correspondence:***
Seth C. Murray
sethmurray@tamu.edu

[†]Present address:
Steven L. Anderson II,
Department of Environmental Horticulture, Institute of Food and Agriculture Science, Mid-Florida Research and Education Center, University of Florida, Apopka, FL, United States

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *23 November 2019*
Accepted: *27 August 2020*
Published: *30 September 2020*

Citation:
Anderson SL and Murray SC (2020) R/UAStools::plotshpcreate: Create Multi-Polygon Shapefiles for Extraction of Research Plot Scale Agriculture Remote Sensing Data. Front. Plant Sci. 11:511768. doi: 10.3389/fpls.2020.511768

Agricultural researchers are embracing remote sensing tools to phenotype and monitor agriculture crops. Specifically, large quantities of data are now being collected on small plot research studies using Unoccupied Aerial Systems (UAS, aka drones), ground systems, or other technologies but data processing and analysis lags behind. One major contributor to current data processing bottlenecks has been the lack of publicly available software tools tailored towards remote sensing of small plots and usability for researchers inexperienced in remote sensing. To address these needs we created plot shapefile maker (R/UAS::plotshpcreate): an open source R function which rapidly creates ESRI polygon shapefiles to the desired dimensions of individual agriculture research plots areas of interest and associates plot specific information. Plotshpcreate was developed to utilize inputs containing experimental design, field orientation, and plot dimensions for easily creating a multi-polygon shapefile of an entire small plot experiment. Output shapefiles are based on the user inputs geolocation of the research field ensuring accurate overlay of polygons often without manual user adjustment. The output shapefile is useful in GIS software to extract plot level data tracing back to the unique IDs of the experimental plots. Plotshpcreate is available on GitHub (https://github.com/andersst91/UAStools).

Keywords: shapefiles, open source software, small plot, agricultural, GIS

INTRODUCTION

Remote sensing platforms geared towards automated high-throughput crop monitoring have become important tools with potential to drive gains in crop improvement and management (Araus et al., 2018). Although curating sensor information/images has become somewhat run-of-the-mill, especially for remote sensing specialists, processing sensor information into informative data for decision making remains a tedious, time consuming, and challenging process (Shakoor et al., 2017; Shakoor et al., 2019). Aside from the processing/calibration of sensor datasets, reducing dataset dimensionality is a critical step in facilitating the ability to make actionable decisions. In plot-based agriculture research programs this requires the creation of individual areas of interest (AOI) for each research entry/treatment of interest. These AOIs are used to extract plot level

information, such as the plant height, canopy cover, or vegetation index of a specific plot containing an individual genotype or experimental treatment. When the number of plots is small (<50), little effort is required and shapefiles containing AOIs can be manually drawn. However, for large plant breeding or genetics programs hundred to multiple thousands of plots, AOI may be needed and unique identifying information with consistent and repeatable labeling is needed for each AOI.

There are several features that are needed to make plot extraction from GIS software efficient, even for novices. (i) The ability to rapidly create a grid of polygons to be overlayed on plots in the proper rotation for any mosaic. (ii) The ability to easily incorporate the experimental design using tabular information with attributes, such as, unique plot IDs for each polygon. (iii) An option for buffering (i.e., a reduced representation of the plot polygon) to exclude areas of bare soil (e.g., walkways/alley) and reduce plot overlap when an orthomosaic has some distortion. (iv) Free and open source availability that allows all researchers to use the same tool without proprietary software.

Tools available to rapidly create AOI polygons for large scale small plot trials (> 100 of plots) are limited (**Table 1**), or unknown to the user community. ArcGIS (ESRI Development Team, 2019) and QGIS (QGIS Development Team, 2019) utilize a fishnet approach to create a regular gridded rectangle, although unique identifiers must manually be assigned to each polygon. Unique ID assignment is further complicated due to the left-to-right, top-to-bottom grid creation rather than the bottom-to-top, serpentine design commonly used in small plot design. ArcGIS and QGIS require identification of a four-point coordinate system to properly orient gridded polygons to the field-plot offset from north-south orientation. Plot Phenix (Progeny Development Team, 2019) "grid" functionality, a commercial software, resolves this issue through manual, point and click identification of corner plots, automated polygon centering, and a vast array of options to optimize polygon size, rotation, buffer, stagger, and subsetting. R/FieldimageR::fieldshape (Matias et al., 2020) and ImageBreed (Morales et al., 2020) "Generate Polygon Template" are open source software that create plot polygons based on manual, point and click identification of polygon grid corners in combination with total column and row counts. R/FieldimageR and ImageBreed link polygons back to unique IDs and plot design, but lack buffering functionality. Additionally, they provide plot, and image rotation capabilities are as separate functions/steps. Although several softwares are beginning to provide automated polygon gridding functions tailored to small agricultural research plots there is still a need for an open source resource that incorporates (i) plot orientation, (ii) experimental design, (iii) automated attribute table with unique plot ID, and (iv) plot buffering.

IMPLEMENTATION

R/UAStools::plotshpcreate (**File S1**) is implemented as a software package function of R (**Figure 1A**), which constructs a multi-polygon shapefile (.shp) of a research trial, with individual polygons defining specific research field plots. Plotshpcreate has two dependency packages (R/sp (Pebesma and Bivand, 2005; Bivand et al., 2013) and R/rgdal (Bivand et al., 2019)) and is recommended to be executed on the most current version of R. Plotshpcreate has three main argument inputs (i) seed preparation and experimental design data frame (**Figure 1B**), (ii) A-B line coordinates (**Figures 1C, D**), and (iii) plot and buffer dimensions (**Figure 1E**). Output files include a multi-polygon ESRI shapefile using overall plot dimension and a multi-polygon ESRI shapefile using buffer plot dimension. Optional outputs include visual representations of shapefile for rapid accuracy assessment. UAStools can be loaded into the R environment using the devtools package (**Figure 1A**). Example scripts can be found using (i) "?plotshpcreate" command in R, (ii) the github wiki page (https://github.com/andersst91/UAStools/wiki/plotshpcreate.R), and (iii) example pipeline scripts (https://github.com/UFResearchComputing/PlantSci_BigData/blob/master/Workshop/UF_PSS_Script_v3.R).

Required Inputs

Seed Preparation and Experimental Design Data Frame

The infile (e.g., R/View(SampleInfile); **Figure 1B**) for plotshpcreate.R requires four columns matching the quoted column names below (additional columns are permitted but won't be utilized): (i) "Plot": The number of each plot (numeric); (ii) "Barcode": A unique identifier for each plot (character); (iii) "Range": The range (also referred to as row in non-furrow irrigated agriculture systems and reflects the rows of your plot design matrix) number of each plot in the plot design matrix (numeric); and (iv) "Row": The row (also called column in non-furrow irrigated agriculture systems) number of each plot in the plot design matrix (numeric). Barcodes must be unique across all observations if nrowplot=1 (i.e., if every observation of the infile has a unique barcode use nrowplot==1). Repeated barcodes and plot numbers if there are multi-row plots as the plotshpcreate

TABLE 1 | Available software which can create gridded multi-polygon shapefiles.

Software	Function name	Connects unique ID	Utilized plot design	Automated buffer zone	Automated Rotation	Open Source
ArcGIS	Create Fishnet	F	F	F	T	F
ImageBreed	Generate Polygon Template	T	T	F	F	T
Plot Phenix	Grid	F	T	T	T	F
QGIS	Vector Grid	F	F	F	T	T
R/FIELDimageR	fieldShape	T	T	F	F	T
R/UAStools	plotshpcreate	T	T	T	T	T

A

```
install.packages("devtools")
library("devtools")
devtools::install_github("andersst91/UAStools")
library("UAStools")
```

B

```
> head(SampleInfile)
   Loc Test Plot       Barcode                      Stock              Pedigree Entry Rep Range Row
1 CS17 G2FE    1 CS17-G2FE-001    16.2.19716.04067.0000000 NILASQ4G31I11S3/LH123HT     1   1     4 255
2 CS17 G2FE    1 CS17-G2FE-001    16.2.19716.04067.0000000 NILASQ4G31I11S3/LH123HT    NA   1     4 256
3 CS17 G2FE    2 CS17-G2FE-002 16.2.19716.04555.0000000.m1           2369/LH123HT    NA   1     4 257
4 CS17 G2FE    2 CS17-G2FE-002 16.2.19716.04555.0000000.m1           2369/LH123HT    NA   1     4 258
```

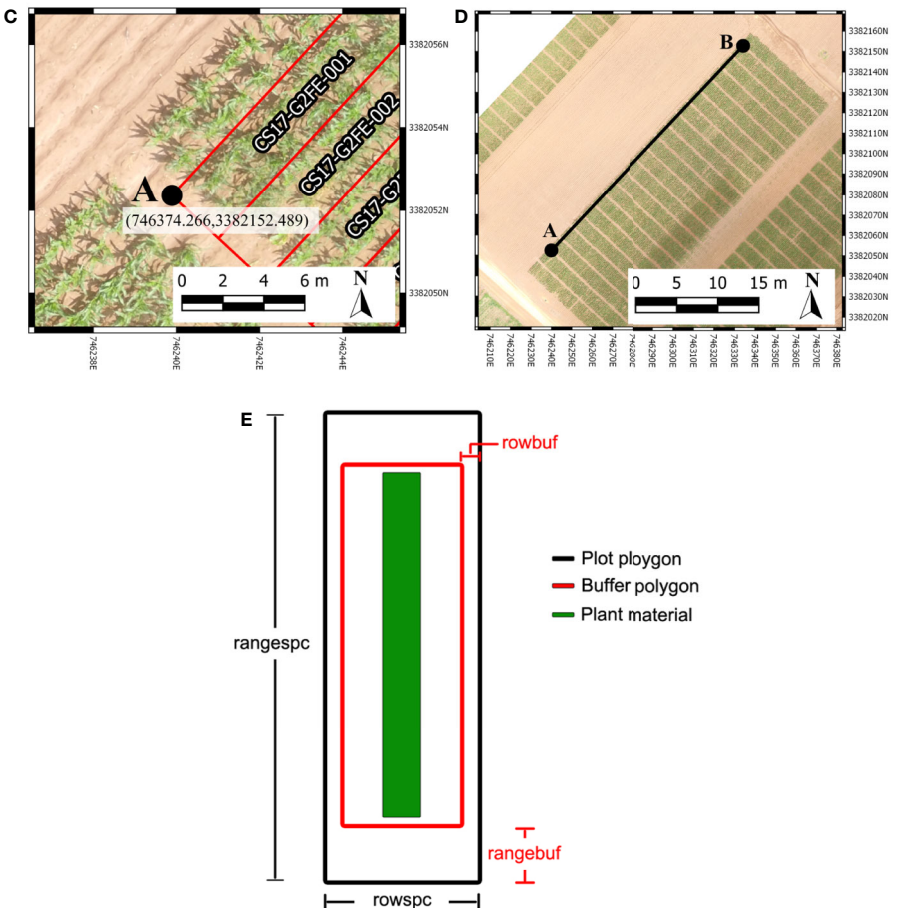

FIGURE 1 | **(A)** Executable lines necessary to load UAStools into the R environment. **(B)** Example of common data structure used as the input file for plotshpcreate. **(C)** Demonstrates the localization of an "A" point in reference to the first plot polygon of respective experimental design matrix (front, left corner of the fist plot if reading a book from bottom to top, left to right). **(D)** Visual representation of A-B line. **(E)** Diagram demonstrating the plot (black) and buffer (red) polygon spacing input parameter. Portions of this figure were adopted from Anderson et al. (2019).

function accounts for this redundancy within the function. Barcodes must be identical across adjacent rows in a plot if trial consists of multi-row plots. An example from the barcode system we typically use is "CS17-G2FE-018" where "CS" denotes the location, "17" denotes the year, "G2FE" denotes a trial in this location year and "018" denotes the 018th plot within this trial. A sample dataset has been provided with R/UAStools and is defined in R as "SampleInfile" when UAStools is loaded *via* library("UAStools") command.

A-B Line Coordinates

Plotshpcreate was developed for Universal Transverse Mercator (UTM) coordinates. Convert to UTM before attempting to use plotshpcreate using a projection transformation tool (i.e.,

R/rgdal::project). Plotshpcreate builds plot polygons based on the "A" point (**Figure 1C**) as a reference and utilized the plot locations in the rectangular grid (Range, Row) of the plot design matrix to calculate the appropriate geo-locations for the polygon corners. The location of "A" is specific, and must lie at the front, left corner of the first plot of respective experimental design matrix (front, left corner of first plot if reading a book from bottom to top, left to right). More specifically, within the middle of the preceding alley and in the middle of the inter row space to the left of the first plot (**Figure 1B**). The B point is less specific but should be place in the same inter-row space to accurately capture the exact angle (i.e., deviation from South/North orientation) of the field (**Figure 1C**).

The best method for the A-B line development is using the geo-rectified orthomosaic of interest, alternatively a high-confidence handheld real-time kinematic (RTK) GPS on a pole to ensure an accurate A-B line in the field. A and B points can be identified using existing R packages (i.e., R/raster::plotRGB along with R/graphics::locator), but we recommend using QGIS (or equivalent software) due to improved resolution for identification of UTM point coordinates. If many temporal orthomosaics will be used throughout the season, one of these with high accuracy (e.g., GCP error) and low distortion can be used to develop plotshpcreate and subsequently applied to all other timepoints. Although use of a single shapefile across multiple orthomosaics is ideal, the user should be aware that error/inconsistencies in image stitching, as well as variances in orthorectification efficiency and accuracy could result in the inaccuracy of shapefile location when used to extract data from orthomosaics other than the reference mosaic. Visual accuracy checks are the simplest way to assess such accuracy issues and identify data sets that may need a dataset-specific shapefile. Use of plotshpcreate.R is possible with orthomosaics lacking geo-rectification, but requires the user to manually identify A and B points within each orthomosaic using a GIS software such as QGIS, ArcGIS, or R/rgdal. Users can loop plotshpcreate.R creating multi-polygon shapefiles for each unique non-georeferenced orthomosaic.

Plot and Buffer Dimensions

There are four polygon dimension arguments that can be specified to accurately create the proper plot dimensions and buffer dimensions desired (**Figure 1E**). Row (i.e., column) spacing (rowspc) spacing of a single row is set to 0.76 m in reference to the row spacing, by default. Range (i.e., row) spacing (rangespc) refers to the total plot length including half alley distance on either side of the plot (default: 7.62 m). Row buffer spacing (rowbuf) is the distance removed from both sides of rowspc to create a buffer zone between plots boundaries (default is 0.03 m). Range buffer spacing (rangebuf) is the distance removed from both sides of rangespc to create a buffer zone between plots boundaries (default is 0.61 m). As an example, if alleys are 1.22 m rangebuf they should be set to 0.61 m to remove 0.61 m from both ends of the polygon. These settings all must be changed for each researchers plots sizes, any default will almost never fit any other research study.

Optional Functionality Arguments

We have designed plotshpcreate to have several useful functionalities that dictate how plot polygons can be created (**Table 2**). Plotshpcreate was developed based on a common style of seed preparation input files, meaning that if a plot consists of multiple planted rows, the input file must contain a row of data for each plot of the design/layout matrix (i.e., every range x row combination) with the same unique ID. There are ways to overcome this by adjusting plot dimensions and input file, but we will not discuss those methods.

TABLE 2 | Gallery of plotshpcreate input parameters.

Parameter	Default	Options	Description
A	NULL	User	Numeric vector of UTM coordinates (Easting, Northing) of "A" point.
B	NULL	User	Numeric vector of UTM coordinates (Easting, Northing) of "B" point.
UTMzone	NULL	User	Character parameter defining UTM zone number. Default will result in an coordinate reference system of "NA".
Hemisphere	"N"	User	Character parameter that designates the Northern "N" or Southern "S" Hemisphere.
infile	NULL	User	Data frame containing seed preparation file and experimental design
outfile	NULL	User	Character assignment to define output file names.
nrowplot	1	>0	Number of adjacent rows that constitute a plot/unique ID
multirowind	FALSE	Logical	Logic parameter that indicates if adjacent plot rows should be combined and treated as a single plot shapefile and unique identifier.
rowspc	2.5	>0	Row (i.e., column) spacing of a single row.
rowbuf	0.1	≥0	Distance removed from both sides of rowspc to create a buffer zone between plot boundaries.
rangespc	25	>0	Range (i.e., row) spacing of a single row.
rangebuf	2	≥0	Distance removed from both sides of rangespc to create a buffer zone between plots boundaries.
stagger	NULL	User	Numeric vector c(i, j, k) of length three defining [i] row where staggers starts, [j] rows sowed by planter in a single pass, and [k] stagger offset distance from A point.
plotsubset	0	≥0	Defines how many adjacent rows should be excluded from either side of the plot.
field	NULL	User	Character vector to indicate the trial the shapefile is being developed for. Example: CS17-G2FE
unit	feet	feet and meter	Character vector that the unit of measure for the polygon dimensions.
SquarePlot	TRUE	Logical	Logic parameter to indicate if PDF file is desired for visualization of non rotated polygons.
RotatePlot	TRUE	Logical	Logic parameter to indicate if PDF file is desired for visualization of rotated polygons.

The default arguments assume single row/range plots (nrowplot=1) and a unique barcode for each row of the input file (equivalent to **Figure 2A**). It is common to have multiple adjacent rows plots where researchers desire a single measurement representing the combined rows. Plotshpcreate combines multi-row plots (**Figure 2B**) based on matching barcodes by defining the number of rows a plot contains (nrowplot="n") and telling plotshpcreate to combine the rows (multirowind=F). Plotshpcreate can create single polygons of each row plot of a multi-row plot (**Figure 2A**), adding an index to each Unique ID in order to identify the data of the multirow plot from left to right (e.g., left row: CS17-G2F-018_1, right row: CS17-G2F-018_2, etc.). Individual row polygons of a multi-row plot can be created with the arguments multirowind=T and defining the number of rows a plot contains (nrowplot="n"). Multirow plots with rows extracted individually in this way can be averaged after extraction or during analysis. However, while a two-row plot (for example) will double the number of observations, these will not be independent, and caution should be used in interpretation of degrees of freedom.

It is common in advanced yield trials to collect data from interior rows of a multi-row plot to factor out neighboring plot competition. Plotshpcreate has a built in sub setting functionally (plotsubset="n") to create polygons of those specific AOIs. The plotsubset argument works by removing "n" rows from either side of the multi-row plot and returns the remaining inter rows and can be used in combination with "multirowind" and "nrowplot" arguments (**Figures 2C, D**). For example, with a six row plot set "plotsubset=2", plotshpcreate will return the inner two rows of the plot removing two rows from adjacent sides. Alternatively, all six individual plots could be extracted and the outer four discarded, however this would result in a threefold

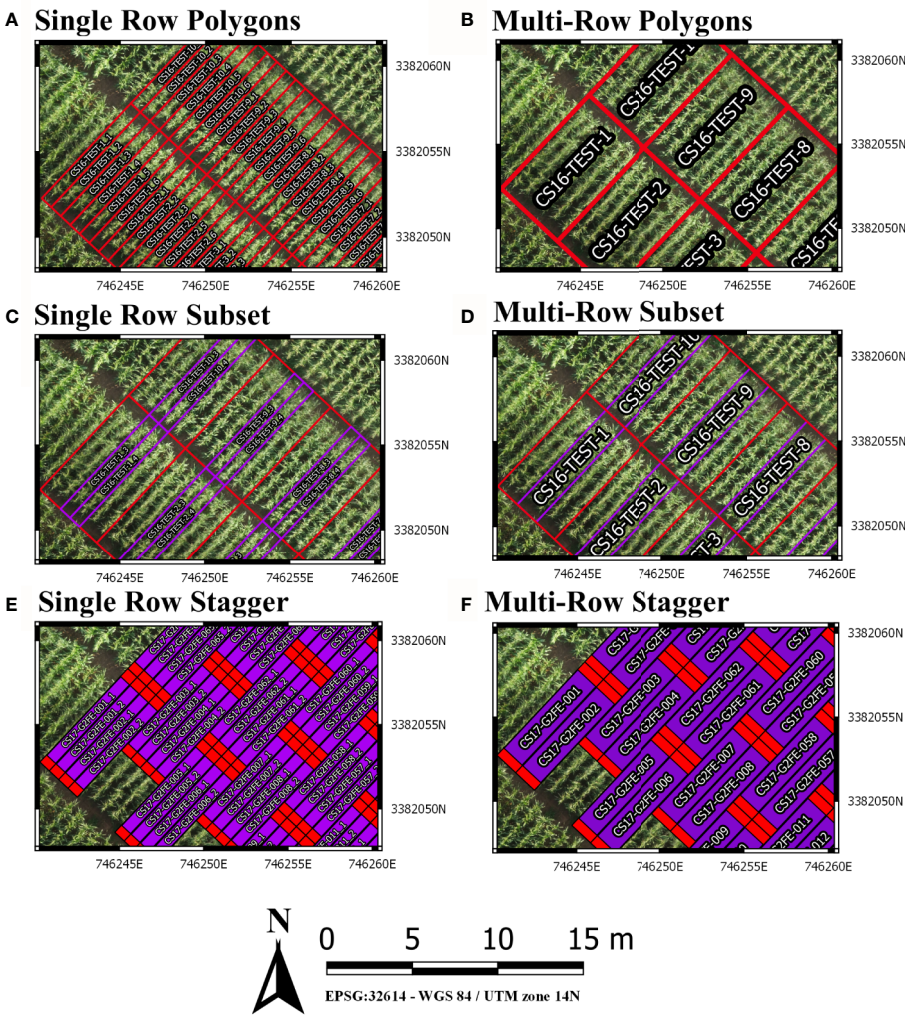

FIGURE 2 | **(A)** Polygons Created for each individual row of a six row plot. **(B)** Single polygons created for each plot merging the adjacent rows of each plot. **(C)** Sub-setting out the middle two rows (purple) of a six row plot (red). **(D)** Sub-Setting out the middle two rows and merging them to a single polygon (purple) of a six row plot (red). **(E)** Staggering individual row plot polygons to adjust for staggered planting. **(F)** Staggering merged two row plot polygons to adjust for staggered planting.

larger file taking additional time to extract and analyze, and the two inner rows would still need to be averaged in some appropriate way.

Furthermore, plotshpcreate can adjust polygon geolocation based on a consistent staggering of plot plantings caused by GIS or tripping issues (**Figures 2E, F**). Plotshpcreate adjusts row and range numbering to begin at one based on the input variable (i.e., if the minimum row number is three it will be adjusted to one, four adjusted to two, and so forth). This is important to remember when utilizing stagger whether field border rows are incorporated within your input file matrix. Plotshpcreate can create staggered plots grids with an input vector (stagger=c(i,j,k)) describing the row where staggers start (i), how many rows the planter sows in a pass (j), and the stagger offset from the "A" point (j). For example, if we set "stagger=c(5,4,3.8)" and include two rows of border plots within the input file, plotshpcreate will create a four row stagger (j), 3.8 m towards the back of the field based on the "A" point (k), beginning at fifth row (i) of the field from left to right (**Figure 3A**). The stagger pattern of the field is based on the planter passes, if you have two rows of border and a four row planter, the stagger would begin on the third row of the trial (e.g., stagger=c(3,4,3.8) if border is not included within your input file (**Figure 3B**). If multirow plots are not spilt across planter passes (i.e., there is not staggered adjacent plot rows) the "plotsubset" and "nrowplot" arguments make be implemented in conjunction with "stagger" (**Figures 2E, F**).

CONCLUSION

Implementation of high throughput phenotyping platforms such as UAS or ground vehicles can provide a vast amount of data rapidly. In contrast, the development of tools to process sensor datasets is in its infancy, or non-existent, and continued development of data analytic tools is critical to aid rapid data analysis for actionable information extraction (Shakoor et al., 2019). As a result, manual data wrangling remains a laborious time sink in processing sensor datasets. Plotshpcreate was developed to overcome a critical time sink within the data processing pipeline, creating AOIs for research plots at scale. Plotshpcreate provides a tool to rapidly create gridded AOI polygons with attached unique IDs for extraction of sensor data on an agriculture research plot scale within seconds, compared to the hours it would require to manually draw polygons and define unique IDs of thousands of plots within a GIS software. Foundational tools, like plotshpcreate, set the basis for developing more advanced point and click graphical user interface tools, such as shiny (Chang et al., 2019). Additionally, incorporating algorithms that utilize the imagery to auto correct

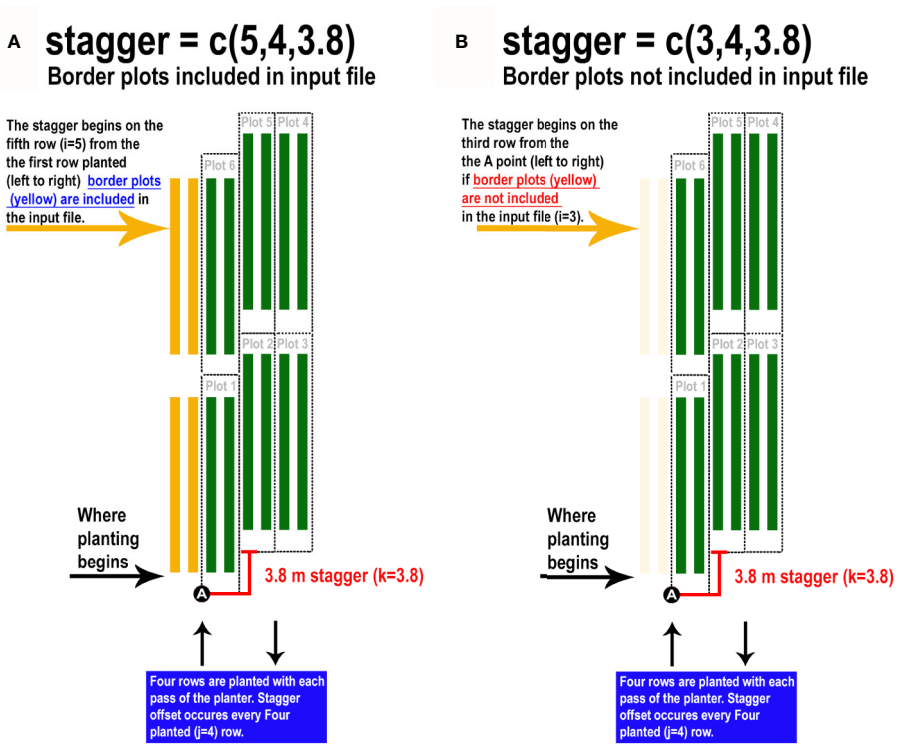

FIGURE 3 | Illustration demonstrating how to properly implement the "stagger" argument of plotshpcreate if **(A)** border plots are incorporated withing the input file design matrix or **(B)** border plots are not incorporated withing the input file design matrix.

for minor changes in plot orientation (Ribera et al., 2017) would be a useful, although it would likely increase computation time and memory with the inclusion of imagery data analysis. Plotshpcreate has room for improvement through increased functionality and the developers encourage the community to aid in adding new tools and they feel necessary.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

Conceptualization: SA and SM. Methodology: SA and SM. Software: SA and SM. Validation: SA. Resources: SM. Data Curation: SA. Writing—Original Draft Preparation: SA. Writing—Review and Editing: SA and SM. Visualization: SA. Supervision: SM. Project Administration: SM. Funding Acquisition: SM.

FUNDING

This research was funded by USDA-NIFA-AFRI Award No. 2017-67013-26185, USDA-NIFA Hatch funds, Texas A&M AgriLife Research, the Texas Corn Producers Board, the Iowa Corn Promotion Board, and the Eugene Butler Endowed Chair in Biotechnology. SA was funded for one year by the Texas A&M College of Agriculture and Life Sciences Tom Slick Senior Graduate Fellowship. The funders had no involvement in the study. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.511768/full#supplementary-material

FILE S1 | R/UAStools v0.2.0 R package.

REFERENCES

Anderson, S. L., Murray, S. C., Malambo, L., Ratcliff, C., Popescu, S., Cope, D., et al. (2019). Prediction of maize grain yield before maturity using improved temporal height estimates of unmanned aerial systems. *Plant Phenome J.* 2 (1), 1–15.

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23 (5), 451–466. doi: 10.1016/j.tplants.2018.02.001

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with {R}. 2nd ed.* (NY: Springer).

Bivand, R., Keitt, T., and Rowlingson, B. (2019). *rgdal: Bindings for the "Geospatial" Data Abstraction Library. R package version 1.4-4.* Available at: https://CRAN.R-project.org/package=rgdal.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). *shiny: Web Application Framework for R. R package version 1.3.2.* Available at: https://CRAN.R-project.org/package=shiny.

ESRI Development Team (2019). *ArcGIS Pro. Environmental Systems Research Institute (ESRI).* Available at: http://resources.arcgis.com/en/help/main/10.2/index.html.

Matias, F. I., Caraza-Harter, M. V., and Endelman, J. B. (2020). FIELDImageR: An R package to analyze orthomosaic images from agricultural field trials. *Plant Phenome J.* 3 (1), e20005. doi: 10.1002/ppj2.20005

Morales, N., Kaczmar, N. S., Santantonio, N., Gore, M. A., Mueller, L. A., and Robbins, K. R. (2020). ImageBreed: Open-access plant breeding web–database for image-based phenotyping. *Plant Phenome J.* 3 (1), e20004. doi: 10.1002/ppj2.20004

Pebesma, E. J., and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R. News* 5 (2), 9–13.

Progeny Drone Development Team (2019). *Phenix*. Available at: https://www.plotphenix.com/.

QGIS Development Team (2019). "QGIS Geographic Information System," in *Open Source Geospatial Foundation Project*. Available at: http://qgis.osgeo.org.

Ribera, J., Chen, Y., Boomsma, C., and Delp, E. J. (2017). Counting plants using deep learning. *2017 IEEE Global Conf. Signal Inf. Process. (GlobalSIP)* 5, 1344–1348. doi: 10.1109/GlobalSIP.2017.8309180

Shakoor, N., Lee, S., and Mockler, T. C. (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol.* 38, 184–192. doi: 10.1016/j.pbi.2017.05.006

Shakoor, N., Northrup, D., Murray, S., and Mockler, T. C. (2019). Big Data Driven Agriculture: Big Data Analytics in Plant Breeding, Genomics, and the Use of Remote Sensing Technologies to Advance Crop Productivity. *Plant Phenome J.* 2 (1), 1–8. doi: 10.2135/tppj2018.12.0009

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership