Heinrich Heine
Universität
Düsseldorf

Assessment of barley genome and transcriptome features and their application in quantitative genetics and plant breeding

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Christopher Arlt
from Dortmund, Germany

Rostock, August 2025

From the institute for Quantitative Genetics
and Genomics of Plants
at the Heinrich Heine University Düsseldorf

Published by the permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Contributors:
1. Prof. Dr. Benjamin Stich
2. Prof. Dr. Karl Köhrer

Date of the oral examination:

Declaration of the Doctoral Dissertation

I herewith declare under oath that this dissertation was the result of my own work without any unauthorized help in compliance with the "Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf". This dissertation has never been submitted in this or similar format to any other institution. I have not previously failed a doctoral examination procedure.

Rostock, 20.08.2025

Christopher Arlt

Contents

---

[1] M. Weisweiler, **C. Arlt**, P.-Y. Wu, D. Van Inghelandt, T. Hartwig, B. Stich. 2022, Theoretical and Applied Genetics, 135:3511–3529

[2] **C. Arlt**, T. Wachtmeister, K. Köhrer, B. Stich. 2023. Plant Biotechnology Journal, 21(11)

[3] **C. Arlt**, D. van Inghelandt, J. Li, B. Stich. 2025, accepted by Theoretical and Applied Genetics

[4] F. Casale, **C. Arlt**, M. Kühl, J. Li, J. Engelhorn, T. Hartwig, B. Stich. 2025, submitted for publication

1 Summary

Plant breeding is an important factor in ensuring the ability of society to react to changes in environmental conditions and to secure the world wide supply with food. Modern plant breeding is an interdisciplinary endeavor. The steady increase in knowledge produced by biological research in plant physiology, reproduction, and genetics enables the development of quantitative genetic methods applied in plant breeding to maximize the success of breeding projects. This thesis combines multiple studies that contribute to this interdisciplinary effort by developing, evaluating, and applying new methods for the globally important cereal crop barley.

The amount of next-generation sequencing (NGS) data available for barley and other crops is constantly increasing. Recent advances in sequencing technology, accompanied by a reduction in sequencing prices for NGS technologies, enabled the scalability necessary in the field of plant breeding. Therefore, NGS data types that were previously only used in small-scale research are now considered for application in a plant breeding context. An example of this is the identification of genome wide structural variation (SV).

SV identification in the large and repetitive genome of barley is difficult, and our study was the first to re-evaluate a broad set of available tools in this context. Six of the most promising SV callers for NGS short read data were selected to determine their individual performance using simulated barley short reads. We identified the best combination of tools for each SV type and achieved a sensitivity of more than 70% and a precision of more than 95% simulating a 25x coverage. We used the best combination of tools to identify SV in 23 spring barley parental inbred lines and found almost 500,000 SV clusters in which the most prevalent type of SV was deletions followed by translocations.

These 23 parental inbred lines were used to create a multi-parent recombinant inbred line (RIL) population using a double round robin (DRR) crossing scheme. The population is made up of 45 sub-populations that each have a unique combination of parents. The DRR population was created, among other things, to investigate the variation of crossing over (CO) events between sub-population. Recombination is, like most biological processes, a regulated system. If information from the parental inbreds could be used to predict the recombination behavior of the offsprings, that would help plant breeders to select the most appropriate parental lines for crossing.

Previous research indicated that methylation and genomic variance affect the

rate and distribution of CO events on a mechanistic level. We therefore used the SV data described above in addition to methylation data for the 23 parents to investigate their impact on meiotic recombination in the DRR population. We found CO hotspots predominantly in proximity to genes with low methylation flanked by large regions of highly methylated DNA. Furthermore, CO was positively correlated with low to medium SV rates. However, SV rates above a certain threshold did not increase the rate at which CO events occurred. We conclude that the recombination landscape is highly predictable in our RIL population on the basis of the parental SV and methylation pattern.

This finding supports breeders in the essential task of choosing the right parental lines for crossing. However, an additional aspect in cross-based breeding projects is the correct selection of crosses in the process of narrowing down candidate genotypes. One of the most costly factors during this process is the cultivation and phenotyping of large populations, and genomic selection (GS) can be used to reduce costs by reducing the number of individuals that are phenotyped. Therefore, cost efficiency is an important factor in GS, and genotyping must be significantly cheaper than phenotyping for GS to be useful.

We had success performing GS using SV and gene expression data in the parental lines, leading us to evaluate the possibility of cost-effective alternatives to SNP based genotyping in the RIL population. Despite the fact that SV showed the largest GS potential in the parental lines, it is not possible to generate SV data based on NGS data for a large offspring population cost effectively. We therefore focused on RNA sequencing (RNA-Seq) and investigated the possibility of generating RNA-Seq data for the progenies in such a manner that would be price-competitive with SNP array data generation.

Our manual library preparation miniaturization approach reduced the consumable costs for RNA-Seq data generation by up to 86.1% compared to the gold standard. Up to 54.5% of these cost savings were realized by miniaturization alone. We rigorously tested the quality of the data produced by the workflow and compared it to libraries without miniaturization, and determined that the quality was not negatively impacted by the miniaturization. However, we found that manual miniaturization is limited by the minimal volume that can be accurately pipetted manually before it leads to an increase in variability between samples.

With this workflow, we successfully established a method for generating RNA-Seq data for a competitive price. Therefore, we continued investigating the GS performance of RNA-Seq data in the DRR population. Similarly to the results for the parental lines, we were able to show for three sub-populations that RNA-Seq data can be successfully used to predict a broad set of agronomic traits. We improved the prediction ability of the RNA-Seq gene expression (GE) and SNP

datasets by quality and function filtering. The combination of RNA-Seq SNP and GE did not result in the expected increase in GP performance. Furthermore, GP performance depended on the trait and was highest when training and testing set were closely related. We achieved the highest prediction ability, exceeding the GP performance of the SNP array, when combining the RNA-Seq SNP with the parental SNPs extracted from the same dataset that was used to identify SV.

In general, we were able to show that a diverse set of additional information for parental genotypes can benefit a modern breeding project at multiple points. We also developed RNA-Seq based GS prediction alternative, creating a more flexible and in certain scenarios more cost-effective way to do GS. This sets the groundwork for further research and development of a new tool set that fully utilizes the advancements in sequencing technology.

2 General Introduction

Plant breeding history and the cultivation of barley

Barley is a crop plant of global agricultural importance (Newton et al., 2011; Visioni et al., 2023). In 2023, barley was the crop with the fifth largest area harvested world wide, after wheat, maize, rice, and soy beans (FAOSTAT). The continuous increase in human population and the increase in per capita crop demand create the need for agriculture expansion (Tilman et al., 2011; Lenaerts et al., 2019). The negative ecological impact due to land clearing is substantial and will continue to increase in the future (Burney et al., 2010; Zabel et al., 2019). This is why improving crop yield is an essential part to ensure future food security (Burgess et al., 2023). Furthermore, climate change will increase the need for drought and heat-stress tolerant crops (Ameer, 2024), for which barley is a promising candidate (Abdelghany et al., 2024).

Domestication and early plant breeding

Around 10.000 year ago, barley (*Hordeum vulgare*) started to be domesticated from its wild relative *Hordeum spontaneum*, which is indicated by archaeological sites in the Fertile Crescent (Badr et al., 2000). Therefore, barley was one of the first crops that was domesticated by humans (Pourkheirandish and Komatsuda, 2007). The first steps of human civilization into agriculture focused on the production and propagation of crops to secure a reliable food source that allowed permanent settlements (Purugganan and Fuller, 2009). Over time, people started to select desirable traits including shorter life cycles, high resistance to biotic and abiotic stresses, and higher food content (Wieczorek, 2012). This process was carried out in many regions of the world independently of each other. Over the span of thousands of years, wild crop species were transformed into many different landraces, which were adapted to regional environments. The landraces were then further domesticated into cultivated varieties, or cultivars, by phenotypic selection. During this time, farmers relied on observation and practical experience alone to produce crop improvements. They were able to fundamentally alter crops without knowing the biological systems in control (Bennett, 2010).

Scientific plant breeding

Scientific plant breeding found first commercial success in the early 20th century after multiple scientific discoveries in the later half of the 19th century. Most importantly, by Mendel describing Mendelian inheritance (Mendel and Bateson, 1925) and Darwin's work on cross- and self-fertilization (Darwin, 1900). The new found knowledge about the genetic makeup of gametes, sexual reproduction, and the mechanisms behind different mating systems enhanced the ability of breeders to create and maintain cultivars and revise the seed production system (Bradshaw, 2017).

The pioneering work of Thomas Hunt Morgan in the field of genetics in the early 20th century was the unification of the chromosomal and Mendelian theory (Morgan et al., 1923). A couple years later, in the 1920s, Hermann Joseph Muller conducted a series of experiments in which he was able to show the mutagenesis of chromosomes. The inheritance of these mutations identified the chromosome as the structure which harbor genes (Muller, 1927, 1928) During this period, it was unclear what a gene was and researchers had different theories. Based on their background, a gene was seen as a unit of inheritance, a unit of recombination, and a unit of mutation, which were not necessarily seen as mutually exclusive (Gayon, 2016). DNA was discovered as hereditary material with the emergence of molecular genetics (Avery et al., 1944). Advancements in the 1950s including the discovery of the structure of DNA by Watson and Crick (Watson and Crick, 1953) and later the decryption of the genetic code (Crick et al., 1957) formed the basis for the modern concept of genes.

All the new knowledge and tools established since the early 20th century applied to the field of plant breeding increasing the ability to modify crop plants. Nevertheless, already early in the 20th century it became clear that not all phenotypes follow the categorical manifestation proposed by Mendel. And after early success in plant breeding by selecting for Mendelian traits, for example many disease resistance genes, it became clear that most agriculturally important traits are polygenically regulated and heavily influenced by environmental factors (Mochida et al., 2020; Zhang et al., 2020). Because of that, these trait show a continuous distribution of trait values. Such traits are termed quantitative traits. The plant response to abiotic stresses and yield traits fall into this category and are of highest importance in future breeding efforts (Friedt and Ordon, 2022).

Early quantitative genetic research developed in parallel to Mendelism from the early to mid 20th century. It was able to use the new found knowledge to create specialized methods to improve the field of plant breeding, leading to some

of the most significant milestones in the 20th century and the Green Revolution (Khush, 2001). The term Green Revolution summarizes a number of key achievements in plant breeding during the 1960s and 1970s, including the development of hybrid maize, high-yielding semi-dwarf wheat, and high-yielding short-statured rice (Hargrove and Cabanilla, 1979; Dalrymple, 1980; Duvick and Cassman, 1999).

This era was marked by international public goods institutions that filled the gap left behind by profit-driven private companies. Institutions such as the International Maize and Wheat Improvement Centre (CIMMYT), the International Rice Research Institute (IRRI) and the Consultative Group on International Agricultural Research (CGIAR) invested in fundamental and diverse plant breeding research to produce public resources and goods without the need to profit from it afterwards (Pingali, 2012). This had a positive global impact by lowering food prices and reducing poverty by increasing grain yield from publicly available crop germplasm.

Since then, advances in molecular biology and biotechnology from the 1990s to today have been heavily focused on the genome. In this post Green Revolution era, plant breeding was changed by high-throughput genotyping, leading to genotype-based selection (marker assisted selection; MAS) and targeted genome modification technologies (e.g. CRISPR-Cas9) (Tahakik et al., 2024). In this era, genetic engineering is the next step in scientific breeding to ensure a continued increase in yield and adaptation of crops to changing environmental conditions. It comes with its own challenges, including the conflict between capitalistic interests of companies and the inequality of access to high-yielding crops by farmers in developing countries. Additionally, new technologies need to be accepted by the public and consequently local governments to find broad application and success (FAO, 2004).

Creating diversity: Meiotic recombination

Sexual reproduction has a multitude of beneficial effects for plants. It enables increased genetic diversity by random shuffling of parental alleles and enhances this effect by meiotic recombination (Peck, 1994). Subsequently, this diversity can be used to adapt to changing environments or to occupy a more beneficial ecological niche (Lei, 2010). Sexual reproduction was also shown to avoid the accumulation of deleterious mutations while promoting even rare beneficial mutations (Peck, 1994). The driving force behind the evolution of all biological organisms is natural selection, as was postulated by Darwin (1859) in year 1859.

The genetic background of natural selection was completely unclear at that time. However, we now know that selection leverages the phenotypic diversity caused by genotypic variation within the population. Genotypic variation is created and maintained within plant populations primarily by mutation and recombination. Mutations and structural variances, large-scale genome alterations between individuals, are discussed in more detail later in the introduction.

Meiotic recombination and transposable elements in barley

The genome is mainly restructured by meiotic recombination and non-homologous relocation, swapping large sections of the chromosomes (Parks et al., 2015). However, many of the large plant genomes are mainly comprised of transposable elements that contributed significantly to genome restructuring (Lee and Kim, 2014). In barley, more than 80% of the genome sequence is derived from transposable elements (Mascher et al., 2017). A total of 350 different families of transposable elements (TE) were identified in barley, most TE belong to the superfamily *Gypsy* or *Copia* which are both long terminal repeat (LTR) retrotransposons (Wicker et al., 2017). TE were shown to accumulate in specific chromosomal regions that can be different between families and change the impact they have on the plant. Retrotransposons are one of the main driving factors behind the increase in the size of plant genomes (Bennetzen and Kellogg, 1997). Further investigation of the most prevalent single TE family *BARE1* showed that recombination can limit the increase in genome size caused by retrotransposons in grass species. Recombination events have the opportunity to separate the active elements of retrotransposons and therefore inactivate them (Vicient et al., 1999).

While this inactivation mechanism is a relevant side effect of meiotic recombination in grass species rich in TE, the major beneficial effect of meiotic recombination is the increase in variability between offsprings to expand the genetic diversity within the population. Increasing knowledge of the mechanisms behind meiotic recombination is important for plant breeding. For example, the ability to increase meiotic recombination could improve the efficiency of backcross schemes (Schuermann et al., 2005).

The molecular biology behind meiotic recombination

The molecular biological pathways leading to meiotic recombination as a result of crossing over are the subject of ongoing research. The process of meiotic recombination starts with a deliberate double-strand break, followed by resection, which creates single-strand DNA by degrading the 5' end on both sides of the double

strand break. This is followed by the strand invasion of the homologous chromosome by the ssDNA overhang, forming a displacement loop structure. Afterward, DNA polymerase synthesizes the previously degraded part of the invading strand. The synthesis incorporates the sequence of the homologous chromosome into the invading strand. After complete synthesis of the invading strand, two pathways are possible.

First, during the double-strand break repair (DSBR) pathway, both chromosomes remain connected to each other and form a double Holliday junction. The Holliday junction is formed by the first 3' overhang of the invading strand after complete synthesis, and the second Holliday junction is formed by the second 3' overhang after binding to the free strand of the homologous chromosome.

Second, during the synthesis-dependent strand annealing (SDSA) pathway, the Holliday structure resolves after the DNA synthesis of the invading strand. The synthesized strand anneals to the 3' overhang of the broken chromosome. The remaining single-strand DNA gap is repaired by ligation, restoring the double-stranded chromosome. During SDSA meiotic recombination is not possible. In case of the DSBR pathway, meiotic recombination is the most likely outcome but depends on how the double Holliday structure is broken up. If on both sides the crossing strands are cut, no crossing over occurs.

Independently from pathway and crossing over event, at least a limited non-reciprocal exchange of genetic information is facilitated. This information transfer is called gene conversion (GC). GC is the process in which two highly homologous but unique sequences are reduced to a single unique sequence after double strand break repair. In the repair process, the acceptor strand is partly degraded and afterwards recreated using the donor strand sequence as template. During meiosis the acceptor and donor strands are provided by two homologous chromosomes, and in the case of an allelic difference, gene conversion is responsible for the removal of unique alleles.

Studies suggest that meiotic recombination pathways and double strand break repair systems are finely regulated and facilitated by multiple similar proteins for which some are involved in multiple of the above mentioned pathways (Chen et al., 2007).

Crossing over frequency and distribution in plants

As described above, not all double-strand repairs result in crossing over events. The average number of crossing over (CO) events per chromosome is four or less for 84% for a variety of eukaryotes, as reviewed by Mercier et al. (2015). The number of DSB during miosis is many times higher (Choi and Henderson, 2015).

Most of the DSB that do not result in CO follow the SDSA pathway (McMahill et al., 2007). Although the recombination rate varies in the intra- and inter-specific context, the range between the minimal and maximal recombination rate appears to be relatively universal (Ritz et al., 2017). A single crossing over event is considered obligatory because both recombination and miosis are mechanistically linked through the formation of chiasmata. Chiasmata, specialized chromatin structures that connect homologous chromosomes during recombination, are formed to allow chromosome segregation during meiosis I (Pollard et al., 2017; Ritz et al., 2017).

The maximum number of COs is primarily limited by the inability of two crossing over events happening near each other. This is called crossing over interference and should space out the crossing over events across the chromosome (von Diezmann and Rog, 2021). Despite this limitation, an increase in CO events could accelerate breeding programs. Understanding the underlying mechanisms of the CO rate and CO distribution is therefore an important goal.

Studies in *Arabidopsis* showed that CO hotspots were associated with low nucleosome density and low DNA methylation regions. CO also increased in proximity to gene promoters (Choi et al., 2013). The positive correlation between gene density and recombination rate in plants was shown multiple times (Paape et al., 2012; Silva and Grattapaglia, 2015; Gion et al., 2016; Wang et al., 2016; Apuli et al., 2020), which would support that the results in *Arabidopsis* are more broadly applicable. Some earlier studies have concluded that CO events occur predominantly in regions with low gene density (Giraut et al., 2011; Yang et al., 2012). The correlation can also change according to the method used, as was shown in Hassan et al. (2021). There is also evidence that epigenetic modification and therefore DNA methylation have a direct impact on the recombination rate (Migicovsky and Kovalchuk, 2013).

Previous research is inconclusive if sequence variance (e.g. SNP) would negatively or positively correlate with CO rate. On the one hand, it is plausible that the variation between homologous sequences hinders the homology search and reduces the binding efficiency during strand inversion (Henderson, 2012). However, multiple studies report a positive correlation between CO rate and SNP density (Saintenac et al., 2011; Salomé et al., 2012; Yang et al., 2012; Bauer et al., 2013; Jordan et al., 2018; Marand et al., 2019). The negative effect of sequence variance on CO could only apply after surpassing a critical threshold, which is supported by studies in *Arabidopsis* (Blackwell et al., 2020; Hsu et al., 2022). An additional aspect which complicates the comparison of most studies in this field is the disparity in the resolution of the underlying genetic map. A low-resolution genetic map is not able capture a complete picture of the recombination landscape and will miss CO events. Many high-resolution studies are conducted in *Arabidopsis* (Lu et al.,

2012; Sun et al., 2012; Yang et al., 2012; Wijnker et al., 2013; Rowan et al., 2019; Fernandes et al., 2024) and some of the most important crops like wheat (Jordan et al., 2018), maize (Li et al., 2015; Rodgers-Melnick et al., 2015), and rice (Si et al., 2015; Marand et al., 2019). For barley, multiple studies investigate meiotic recombination on a broad genomic scale (Higgins et al., 2012; Dreissig et al., 2019, 2020; Casale et al., 2022).

This thesis tried to add to the current research on this topic. The contradictory results of previous studies regarding the correlation of CO and genomic features creates the necessity to further investigate this association at the species level. Furthermore, because most reported recombination rates in barley are based on coarse genomic maps, the high-resolution genetic map created in this study will be beneficial for future research evaluating the complete genetic variation generated during meiosis.

Creating diversity: Genomic variation

Mutations are a natural occurrence in all biological organisms and ensure variability which is leveraged by evolutionary processes to increase the fitness of the population over time. Mutations can occur spontaneously in the cell caused by DNA oxidation or are introduced by errors during DNA replication and DNA repair. Mutations can also be caused by external factors, called mutagens, which can be chemicals or radiation (Brown, 2002). The mutation rate is limited by DNA repair processes and can vary greatly between species (Danovi, 2023).

Mutation types and their effects

Mutations can be categorized by size and effect on the genome sequence. Large-scale mutations include chromosomal deletions, duplications, translocations, and inversions. Duplications can be further categorized into interspersed duplications and tandem duplications, which can lead to copy number variation (CNV) (Griffiths et al., 2020). All mutations that include more than 50 bp are categorized as structural variance (SV) (Mahmoud et al., 2019).

Small-scale mutations include insertion, deletion, or substitution of one or a few nucleotides (Griffiths et al., 2020). Mutations comprised of more than one but less than 50 bp are classified as small insertions and deletions (InDels). If only one nucleotide is affected, the alteration is classified as point mutation. Point mutations occurring in the coding region of the genome can be further classified on the basis of their effect on the protein sequence (Brown, 2011). Insertions and

deletions lead to frame-shift mutations by changing the triplet structure of all downstream codons. A nucleotide substitution can lead to a synonymous substitution or nonsynonymous substitution. In case of a synonymous substitution, the substitution did not change the amino acid because of the redundancy within the genetic code. Non-synonymous substitution can lead to a missense mutation in which one amino acid is switched for another amino acid, or a non-sense mutation in which the mutation caused a premature stop signal. Point mutations in the non-coding region of the genome are considered neutral in this context. Additionally, mutations can have different effects on the genes in which they are present, from inactivation of the gene function (loss-of-function mutation) or the enhancements of the gene function (gain-of-function mutation) (Brown, 2011).

Structural variation detection

The detection of structural variants can be challenging depending on the type of variation, the complexity of the genome, and the available data. Because SNP data can also infer additional genomic features beyond point mutations, it was used to detect structural variation, mainly copy number variations, but was limited by the available marker data (Bickhart and Liu, 2014). This is why SV detection nowadays is mostly based on sequencing data.

Fundamentally, SV are identified by comparing two DNA sequences. A complete and genome-wide characterization can therefore only be achieved by analyzing whole genome sequencing data, ideally two complete chromosome scale genome assemblies. Creating a sufficiently complete assembly for large repetitive plant genomes is difficult and, more importantly, prohibitively expensive (Claros et al., 2012). This method is therefore almost never used if the primary goal is structural variation detection. This is especially true for barley, considering that the first chromosome scale assembly was created with considerable effort in 2012 (Mayer et al., 2012). Because of that, complete and reliable chromosome scale genome assemblies are still rare. However, a single reference sequence is available for many species that can be used to facilitate SV detection.

The reference can be used to map the sequencing output directly and infer the genome-wide structural variation compared to the reference sequence. In this scenario, compared to the use of two assemblies, the detection of structural variance is much more difficult due to multiple factors (Mahmoud et al., 2019). First, depending on the sequencing data available, the structural variance could be larger than the read itself, and therefore large structural variances cannot be characterized by a single read in isolation (e.g. assembly of the larger structure of the genome around the SV). Second, sequencing errors can obscure the real

underlying SV pattern. Third, structural variance can overlap and nest within each other, which is a general difficulty but is even more pronounced when looking at a narrow context. Due to the difficulties mentioned above, it is preferable to use long read data for SV calling (Jayakodi et al., 2024). However, the creation of long read data is still more expensive than short read data and in a highly competitive field, like plant breeding, the most cost-efficient data type has the biggest chance to be integrated into a breeding project. From the above-described challenges arises the need for a thorough investigation of the available tools to use short read data for SV calling to identify the best workflow for each type of SV.

Structural variation in quantitative genetics

Based on the early success of GWAS and linkage mapping, the association between agronomic traits and large regions of the genome indicated that SV could be very useful as a predictor of quantitative traits (Yuan et al., 2021).

Previous research also suggests that SV may be more likely to have a causal relationship with phenotypic traits than SNP, which was shown in human research multiple times (Alkan et al., 2011; Baker, 2012; Sudmant et al., 2015; Schule et al., 2017; McColgan and Tabrizi, 2018). In plant research, the causal relationship between phenotype and SV was shown many times. In maize, a SV was strongly correlated with alumininum tolerance (Maron et al., 2013). Xu et al. (2012) showed that a SV in rice plays a central role in domestication and disease resistance. In wheat SVs were identified to correspond to plant height (Li et al., 2012) and heading date (Nishida et al., 2013). Research in barley shows an association between boron toxicity tolerance (Sutton et al., 2007) and disease resistance (Munoz-Amatriain et al., 2013) and SV. In recent years, many more studies evaluated SV in a pan-genome context by utilizing NGS data. For example, the influence of SV on disease resistance in wheat (Walkowiak et al., 2020), *Brassica napus* (Dolatabadian et al., 2020), and *Brassica oleracea* (Golicz et al., 2016; Bayer et al., 2019), flavor and fruit size in tomato (Gao et al., 2019), and stress response genes in rice (Fuentes et al., 2019). Alonge et al. (2020) was able to detect haplotypes in potato that were not found in previous GWAS using SNP marker data. This indicates that SV is able to identify unique associations and is best used complementary to SNP based analysis. Comparing the ability of SV and SNP to predict trait data as part of a genomic prediction cross-validation scheme can further illuminate the unique benefits of using SV to assess quantitative traits.

This thesis tried to add to the current research on this topic. The only data available from a genome-wide SV study in barley focused on large SV. Therefore, our comprehensive set of SV in 23 diverse spring barley accessions is a beneficial

resource for future barley research. The role of SV in gene regulatory mechanisms and the ability of SV to predict quantitative traits was examined, both of which are mostly unexplored topics in small grain cereals.

Methods in quantitative genetics and plant breeding

One of the most significant points of convergence between Mendelian genetics and quantitative genetic research are point mutations. Point mutations, in this context often described as single nucleotide variation (SNP), can be beneficial or detrimental to the organism (Kono et al., 2016). While the beneficial mutations become fixed within the population, detrimental mutations are generally excluded from the gene pool (deleterious mutations), both through the process of natural selection (Kono et al., 2016). In Mendelian genetics, the search for direct and causal links between these mutations and phenotypes was a primary goal in the last decades. As mentioned above, plant breeding benefited from these findings by facilitating the improvement of Mendelian inherited agronomical traits. For quantitative traits on the other hand, no single gene regulates the target traits and it was therefore impossible to create a link in the same fashion. Therefore, in the field of quantitative genetics, these mutations were seen from a different perspective. SNPs were recognized as genomic markers that indicate the genetic background of the genome surrounding them. Many such markers throughout the genome create a genetic map that captures the broad genomic structure.

Multiple alternative marker types were used before the SNP marker became the standard marker type in plant breeding, including simple sequence repeats (SSRs or microsatellites) and restriction fragment length polymorphisms (RFLP). The first SNP microarray was developed in 1996 for humans by Affymetrix (USA). It captured 1494 SNPs and similar arrays were quickly created for the purpose of plant breeding research.

SNP arrays allowed for excellent scalability and therefore large-scale implementation. One of the first methods that leveraged the scalability was genome-wide association studies (GWAS). The goal of GWAS was to find significant associations between genetic variations and a target trait in a large number of diverse individuals. The region in the genome that shows significant association is called quantitative trait locus (QTL). GWAS in its simplest form is a statistical test for a correlation between the marker allele and the trait expression at each marker position to identify QTL.

The advent of SNP arrays also affected related methods that previously used

different markers such as microsatellites, RFLPs, and transposable element positions (Miles and Wayne, 2008). The advantage of SNP compared to SSR is the frequency in which they occur in plants. SNP based analysis therefore profited from a higher density of markers which was more or less necessary when comparing a highly diverse set of individuals in GWAS, but was also highly beneficial in linkage mapping.

In contrast to association mapping in GWAS, linkage mapping does not rely on natural occurring genetic variation within individuals to find QTL, but on the recombination of parental polymorphic marker. The mapping population is limited to two alleles and the linkage between the markers is dictated by the recombination, leading to slower LD decay compared to association mapping. The observed link between trait and DNA marker, in the form of QTL, allowed breeders to leverage marker-assisted selection (MAS) (Fernando and Grossman, 1989; Lande and Thompson, 1990; Dekkers and Hospital, 2002). In many scenarios where the phenotype is difficult and/or expensive to assess, it is more efficient for breeders to use marker data to select candidates (Budhlakoti et al., 2022). SNP arrays enabled fast and cost-efficient genotyping of large populations. The number of species for which the SNP array is available increased over the years, as did the marker density (e.g., the number of markers per collection).

MAS is most effective when large effect loci could be identified (Heffner et al., 2009). At the same time, most of the markers did not show a major marker trait associations and were therefore not considered by the breeder. Additionally, agriculturally relevant traits are usually traits controlled by many genes with only a small effect (Mackay, 2001; Glazier et al., 2002). To be able to better predict complex quantitative traits with many small-effect loci (Zhao et al., 2014), genomic selection (GS) was developed (Meuwissen et al., 2001). In contrast to MAS, GS estimates breeding values based on the complete set of available marker data. This shift towards GS was supported by the increasing availability of high-density genome-wide marker data.

Genomic selection: Current state and recent advancements

The concept behind genomic selection

In genomic selection, a differentiation is made between training set and a testing population. Most often, both are subsets of a single breeding population or are in

other ways closely related (Alemu et al., 2024). For the training set, both genotypic and phenotypic information is available, while for the testing set, only genotypic data is collected.

The first step in genomic selection is to train a predictive model using the genotypic and phenotypic data of the training set. This model can be a linear regression model, Bayesian model or non-parametric model. A method that is often applied in plant breeding uses a linear regression model based on best linear unbiased prediction (BLUP) to predict the trait (Henderson, 1984). A variant of the BLUP method, genomic best linear unbiased prediction (gBLUP), utilizes a variance-covariance matrix that represents the genetic differences between individuals based on marker data (Clark and Van Der Werf, 2013). In gBLUP, the variance-covariance matrix, also called the genetic relationship matrix, is traditionally created using SNP marker data. However, any type of data describing an heritable relationship between individuals in a population could be used to create a similar matrix. Next, the model is used to estimate the trait values for the testing set based only on the genotypic information (Merrick and Carter, 2021). Lastly, the trait value estimates can be used to calculate the genetically estimated breeding value (GEBV). GEBV describes the genetic potential for a heritable trait and is helpful for the breeder to select the most promising individuals in the breeding population for further propagation.

To evaluate the prediction ability in a research context, sets of individuals for which both genotypic and phenotypic information is available are artificially segmented into testing and training set to simulate the scenario described above. The method most commonly used for these validation runs is the k-fold cross validation. K-fold cross-validation splits the population into k equal sized parts. For example, 5-fold cross-validation divides the entire population into 5 parts with 20% of the data included in each part. For the individuals of one part, the phenotypic data is removed, creating the testing set. The remaining parts build the training set which is then used to predict the testing set trait values. Finally, the predicted trait values are correlated with the removed trait data to estimate the prediction ability for the applied model.

If the predictions are accurate, the breeder is able to potentially reduce costs by decreasing the number of individuals that had to be phenotyped and the number of plants that had to be planted in the field (Wang, 2023). Individuals in the test set can be genotypes based on a small amount of young plant material, which means that cultivation is much more time and space efficient.

The broadening and evolution of GS: The transcriptome

Advancements in sequencing technology brought new alternatives to the SNP marker (e.g., sequencing genotyping) that made GS more broadly applicable (Bhat et al., 2016). But sequencing technology not only creates new types of genomic datasets, it also opens up the possibility for non genomic data sets, for example the transcriptome. Additionally, non-sequence datasets such as the metabolome can be used for GS. Both the transcriptome and the metabolome represent intermediate regulatory stages between genome and trait manifestation and therefore have unique potential to add to the ability to predict the trait using genomic selection. The transcriptome can be quantified using a microarray or mRNA sequencing (RNA-Seq). RNA microarrays are less versatile than RNA-Seq data, for example, because RNA-Seq data also include sequence information from which sequence variant data can be extracted. Both SNP and InDels found in the transcribed portion of the genome, as well as the normalized gene expression values, can be used to create the genetic relationship matrix (Azodi et al., 2020; Weisweiler et al., 2019).

Several plant breeding studies suggest that including transcriptome and metabolome (multi-comics) data in GS has the potential to increase prediction abilities. Combining transcriptomic, metabolomic, and genomic markers data increase prediction ability in diverse maize inbred lines (Guo et al., 2016). Hu et al. (2021) used transcriptomic and metabolomic data to compare single-environmental trails with multi-environmental trails in oat. In addition, a multi-omics study in wheat found that adding incomplete RNA-Seq data to complete genomic marker data helped to assess disease resistance phenotypes Michel et al. (2021). Only recently have studies emerged using multiomics GS in barley (Wu et al., 2022). Most available research uses diversity panels to test the GS potential (Hu et al., 2021; Michel et al., 2021; Westhues et al., 2019; Schrag et al., 2018; Westhues et al., 2017; Guo et al., 2016). And yet, most breeding projects center around large half-sib or full-sib populations for which the GS potential has not yet been validated.

Besides the multi-omics approach to GS there are multiple additional advancements in thisfield. For example, the prediction ability of multivariate GS models could be increased by including high-throughput phenotyping data as a predictor (Rutkoski et al., 2016). Furthermore, multi-trait GS models (Tsai et al., 2020; Lyra et al., 2017; Jia and Jannink, 2012) and multi-environmental GS models (Hu et al., 2023; Li et al., 2019) expanded on the traditional GS framework. Most recently, machine learning algorithms have been leveraged to further increase the performance of GS models (Sandhu et al., 2021; Montesinos-Lopez et al., 2021; Bayer et al., 2021; Washburn et al., 2020; Harfouche et al., 2019).

This thesis tried to add to the current research on this topic. The amount of multi-omics GS research in barley is very limited and to the best of our knowledge,

no previous research investigated the multi-omics GS for a breeding population. Additionally, most multi-omics studies do not fully utilize the RNA-Seq data. Our investigation of the prediction ability potential of RNA-Seq data of three partially related spring barley RIL populations, which include functional parameter filtering and quality filtering, is therefore highly relevant.

RNA-Seq in the context of plant breeding

Cost-efficient creation of NGS data

Next-generation sequencing (NGS) sequencing has become an integral part of many projects related to genetics, genomics, and transcriptomics based on continuous advances in the field over the last 20+ years (Elaine, 2011; McCombie et al., 2019). However, in competitive fields such as plant breeding, multiple factors hinder the broad adoption, despite the fact that sequencing costs are steadily decreasing (Kris, 2021).

In European spring barley breeding for example, a broad research background on a large spring barley accession diversity set laid the foundation for a robust and good performing 50k SNP array (Bayer et al., 2017), which is now well established and has proven itself to be useful in breeding. It is also available for only a fraction of the price compared to competing NGS based datasets. Every NGS workflow has to therefore compete in price or present advantages which would justify the price increase. The price gap between WGS and SNP array genotyping is too large to be bridged at the moment. Therefore, multiple NGS-based genotyping tools were developed to reduce complexity and costs, for example, reduced-representation library (RRL), restriction-site-associated DNA sequencing (RAD-seq), and Genotyping by sequencing (GBS) .

The need for complexity reduction is greatest in plant species that have comparatively large genomes, including barley. In contrast to the highly variable genome size of plant genomes, the size of the transcriptome is relatively stable and much smaller. In barley, the exome is 61.6 Mbp (Hisano et al., 2017), which is only around 1.2% of the estimated whole genome size (Mascher et al., 2017). Additionally, the transcriptome is a biologically highly relevant sequence subset and therefore an interesting alternative approach to reduce costs of NGS-based genotyping.

The most cost-effective way to generate short read sequencing data is on the largest sequencing platforms which generate more than 1000 Gbp of data per run. For reduced complexity sequence datasets like RNA-Seq, this amount of

sequencing data is sufficient for multiple samples leading to the next cost-saving approach: multiplexing. The principle behind multiplexing is to reduce costs by combining multiple samples prior to sequencing. Each sample within these sample pools is uniquely labeled and can be separated after sequencing. The number of samples in each pool can vary and is usually limited only by the combination of unique bar codes available. This can be used to optimize the sequencing data output per sample and aim for the highest possible cost reduction based on the project requirements.

However, sequencing data are only one of two main contributors to the overall costs of RNA-Seq data. The second contributor is sample preparation, and here mainly library preparation. Based on the above-described concepts, the sequencing costs itself can be greatly reduced assuming a equivariant cost reduction is possible for the library preparation enabling the generation of RNA-Seq data at lower costs than SNP Array data. To further maximize the cost-saving potential, the sequence library complexity is further reduced by 3' or 5' end enriched RNA-Seq (Joseph et al., 2019; Luisa et al., 2019; Jeffery et al., 2019; Evan et al., 2015). Traditionally, all transcribed mRNA sequences are used to create library fragments. After random fragmentation of the transcripts, the library fragments are created and sequenced. Finally, sequencing reads can be used to recreate full-length mRNA sequences. The end enriched RNA-Seq approach creates library fragments only from the end pieces of the transcript, most often the 3' end of the transcript which can be easily captured based on the poly-A tail. This removes the possibility of recreating the sequence of complete mRNA molecules, but the ability to assess the relative abundance of the transcript is retained.

Reducing library preparation costs

The pressure to reduce costs increases with the number of samples. This challenge emerges in more and more research fields, for example, in single cell RNA-Seq, but also in plant breeding because most quantitative genetic analysis relies on a high sample count to ensure statistically meaningful results. The above-described multiplexing is able to reduce the costs of sequencing but traditionally does not reduce the library preparation costs because the attachment of the unique bar code is added towards the end of the library preparation process. Moving the bar coding to the start of the library preparation process means that the sample pool can be created early (early multiplexing), and the number of reactions for all follow-up steps are significantly reduced. It therefore has the potential to significantly reduce material and labor costs and the earlier the multiplexing, the higher the benefit.

Early multiplexing is often combined with 3' end-enriched RNA-Seq to max-

imize cost reduction (Daniel et al., 2019; Johannes et al., 2018; Tamar et al., 2016; Soumillon et al., 2014). This type of workflow modification was initially published as customized protocols and required significant time investment to be implemented. Some became available as commercial services since the initial publication. The service can be more easily integrated into a breeding project, but the potential for cost savings is reduced compared to the independent implementation of the customized protocol.

However, in plant breeding, genomic variation is traditionally used as a marker in various quantitative genetic analyzes, the fact that RNA-Seq can provide sequence variance data for the expressed part of the genome on top of the gene expression data is highly valuable. Therefore, the end-enriched RNA-Seq approach is potentially less suitable in this context.

Library miniaturization

An additional approach to reduce sample preparation costs is library miniaturization. Miniaturization describes the systematic reduction of all reaction volumes that are part of the library preparation workflow. This leads to a significant reduction in material costs and can be combined with automatization, which further reduces costs by reducing hands-on time. Library miniaturization is easy to implement using commercially available library preparation kits, which makes it an achievable alternative to standard in-house sample preparation. Therefore, the potential cost savings would not be reduced by commissioning the sample preparation.

Most studies evaluating the miniaturization potential of commercial library preparation protocols use advanced liquid handling automatization to achieve high reproducibility when transferring small volumes (Baptiste et al., 2020; Samuel et al., 2020; Madeline et al., 2019; Sergio et al., 2016). The miniaturization factor is determinant by technical limitations, for example, minimal transfer volumes, and can be as high as 1:10, which reduces the material costs for library preparation to 10% of the initial costs. At the same time, the liquid handling robots reduce the hands-on time and therefore labor costs. However, initial investments in acquiring laboratory equipment to automate library preparation are in most cases prohibitively expensive. Even if the initial investment was possible, it would offset any potential cost savings for a long time.

An attractive alternative could be manual miniaturization, which would enable the miniaturization of library preparation without the need for additional tools. Similarly to the reaction volume, the RNA input and the library output of miniaturized library preparation are equally effected by the miniaturization

factor. Although the reduced input amount can have a negative impact on the library quality by reducing the library complexity, the reduced output is less important. In most situations, the reduced output is sufficient for at least one sequencing run, and additionally, the required amount per sample is reduced when pooling sequencing samples.

Data quality: Library complexity and biases

Library complexity, the final number of unique molecules per library in solution, is an important quality metric for next-generation sequencing libraries. It represents the potential of the given library to produce a complete picture of the genome or transcriptome. Reduced library complexity can be an indicator of problems during library preparation or insufficient sequencing depth (Rochette et al., 2023).

At multiple points of the library preparation, the loss of unique molecules is inevitable, based on imperfect recovery rates during mRNA capture and magnetic bead washing. Additionally, deliberate exclusion of fragments outside the required size range and incomplete conversion rate during cDNA synthesis. Of cause, each expressed transcript is present multiple times per cell and, assuming that the starting RNA was collected from potentially hundreds of thousands of cells, the loss of fragments during library preparation does not necessarily mean a noticeable reduction in library complexity. Nevertheless, it is important to note that lowly expressed transcripts are more likely to be lost during library preparation.

Most often the library complexity is measured based on the sequencing results and depends therefore also on the sequencing depth, assuming that not all unique fragments are always sequenced. This is in fact most relevant when it comes to sequencing the transcriptome cost efficiently. Most highly expressed genes are easily captured, while less abundant transcripts can only be detected by increasing the sequencing depths. Increasing sensitivity by increasing sequencing depth inevitably leads to higher duplication rates and an excess of reads from highly expressed transcripts, making sequencing less cost-effective.

The number of fragments that have to be sequenced to capture a specific proportion of all unique fragments depends not only on the number of unique fragments in the library, but also on the average duplication rate of these fragments in the library, which is determined by the amount of input material and the PCR amplification. Most NGS library protocols include a PCR amplification step that duplicates available library fragments based on the number of amplification cycles. These interactions illustrate the importance of balancing the amount of RNA input, PCR amplification, and sequencing depth to optimize the library complexity and, therefore, the amount of unique data produced.

Measuring library complexity

Miniaturization not only reduces the input amount of RNA, but also leads to an increase in sample-to-sample variation due to more inaccuracies when pipetting smaller volumes. This makes library complexity an important quality metric for evaluating the library preparation workflow and is mainly measured by the following two metrics. First, the read pair duplication rate is higher for libraries with lower library complexity (Adriana et al., 2014). The reduced number of unique molecules increases the chance that multiple identical library molecules are sequenced. Second, the number of detected genes is smaller when the library complexity is low (Elisabetta et al., 2020). Less abundant transcripts can be lost during library predation and therefore can not be detected afterwards, resulting in a lower number of detected transcripts. The comparison is most useful when comparing samples that were created using the same workflow because the library preparation kit itself was shown to have an effect on library complexity (Louise et al., 2016). Less RNA input can lead to reduced library complexity as well (Adriana et al., 2014). The same was also shown for the DNA input amount in WGS experiments Samantha et al. (2020).

Potential library biases, on the other hand, can be introduced as early as the RNA isolation step. RNA degradation can lead to biased expression levels, as Irene Gallego et al. (2014) showed by testing samples with various RNA integrity Number (RIN) values. During library preparation itself, the best documented step to introduce a bias is the PCR amplification. There, fragment length and GC biases can be introduced depending on the polymerase used as described by Jesse and Matthias (2012), which can also impact differential gene expression (DGE) (Wei et al., 2011). Related biases are related to the sequencing method. Coverage and error biases vary between different sequencing platforms and are introduced by the impact of different base composition on the polymerase during sequencing (Michael et al., 2013).

This thesis tried to add to the current research on this topic. To the best of our knowledge, almost no research has been published that describes manual library miniaturization. Furthermore, no comprehensive evaluation of the impact of manual library miniaturization on library complexity and library biases is available. It is therefore highly beneficial to provide a template for future recreation of the miniaturization method and also include the necessary data to evaluate the effects of the miniaturization on the resulting sequencing data.

Objectives of this thesis

The objective of my thesis was to use advancements in sequencing technology to create genomic and transcriptomic resources to improve future breeding efforts in barley. I wanted to achieve this by exploring innovative ways to use structural variance, methylation, and transcriptome data in a quantitative genetic context.

In particular, the objectives were to:

1. benchmark multiple SV calling workflows and assess the sensitivity and precision to detect SV in the barley genome,
2. characterize the type and distribution of SV cluster in 23 barley inbreds to facilitate quantitative traits mapping in the double round robin population,
3. investigate the association between SV clusters and transcript abundance,
4. and evaluate the ability to predict quantitative traits using SV cluster;
5. develop an easy-to-reproduce, manual miniaturized full-length mRNA sequencing library preparation workflow to facilitate RNA sequencing data generation in the context of large breeding populations
6. and investigated the capabilities of manual miniaturization by evaluating their effect on library complexity and library bias;
7. explore the capabilities of low-cost RNA-Seq data to perform genomic prediction in three barley RIL populations for 8 quantitative traits,
8. optimize the prediction ability potential by combining genomic and transcriptomic data, functional parameter filtering, and empirical quality filtering,
9. and examined multiple additional optimization parameters that lead to cost and time savings;
10. identify the genomic features that best explain recombination in the double round robin population,
11. create a high resolution recombination map for that population,
12. and analyze the association of SV and methylation with recombination hotspots and coldspots.

References

Abdelghany, A. M., Lamlom, S. F., and Naser, M. (2024). Dissecting the resilience of barley genotypes under multiple adverse environmental conditions. BMC Plant Biology, 24(1).

Adriana, A., Caroline, B., Stéfan, E., Laurie, B., Céline, O., Laura, B., Corinne, C., Laurène, G., Corinne Da, S., Cyril, F., Jean-Marc Marc, A., and Patrick, W. (2014). Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. BMC Genomics, 15(1):912.

Alemu, A., Åstrand, J., Montesinos-López, O. A., Isidro y Sánchez, J., Fernández-Gónzalez, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J., Ortiz, R., and Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. Molecular Plant, 17(4):552–578.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. Nat Rev Genet, 12(5):363–76.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., Aganezov, S., Ranallo-Benavidez, T. R., Lemmon, Z. H., Kim, J., Robitaille, G., Kramer, M., Goodwin, S., McCombie, W. R., Hutton, S., Van Eck, J., Gillis, J., Eshed, Y., Sedlazeck, F. J., van der Knaap, E., Schatz, M. C., and Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell, 182(1):145–161 e23.

Ameer, M. (2024). Drought-resistant crops: A sustainable solution to climate change. Int. Res. J. Plant. Sci.

Apuli, R. P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., Street, N. R., and Ingvarsson, P. K. (2020). Inferring the genomic landscape of recombination rate variation in european aspen ( ). G3-Genes Genomes Genetics, 10(1):299–309.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types :

Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. J Exp Med, 79(2):137–58.

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S. H. (2020). Transcriptome-based prediction of complex traits in maize. Plant Cell, 32(1):139–151.

Badr, A., M, K., Sch, R., Rabey, H. E., Effgen, S., Ibrahim, H. H., Pozzi, C., Rohde, W., and Salamini, F. (2000). On the origin and domestication history of barley (hordeum vulgare). Molecular Biology and Evolution, 17(4):499–510.

Baker, M. (2012). Structural variation: the genome's hidden architecture. Nat Methods, 9(2):133–7.

Baptiste, N. J., Emilio, Y., Lorenzo, G., Annina, D.-L., Merit, K., Theofanis, K., and Sebastian, J. (2020). Miniaturization of smart-seq2 for single-cell and single-nucleus rna sequencing. STAR Protocols, 1(2):100081.

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Rincent, R., Schipprack, W., Altmann, T., Flament, P., Melchinger, A. E., Menz, M., Moreno-González, J., Ouzunova, M., Revilla, P., Charcosset, A., Martin, O. C., and Schön, C. C. (2013). Intraspecific variation of recombination rate in maize. Genome Biology, 14(9).

Bayer, M. M., Rapazote-Flores, P., Ganal, M., Hedley, P. E., Macaulay, M., Plieske, J., Ramsay, L., Russell, J., Shaw, P. D., Thomas, W., and Waugh, R. (2017). Development and evaluation of a barley 50k iselect snp array. Frontiers in Plant Science, 8.

Bayer, P. E., Golicz, A. A., Tirnaz, S., Chan, C. K. K., Edwards, D., and Batle, J. (2019). Variation in abundance of predicted resistance genes in the pangenome. Plant Biotechnology Journal, 17(4):789–800.

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., and Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. Plant Genome, 14(3).

Bennett, A. B. (2010). A plant breeder's history of the world. Science, 329(5990):391–392.

Bennetzen, J. L. and Kellogg, E. A. (1997). Do plants have a one-way ticket to genomic obesity? Plant Cell, 9(9):1509–1514.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., and Prabhu, K. V. (2016). Genomic selection in the of next generation sequencing for complex traits in plant breeding. Frontiers in Genetics, 7.

Bickhart, D. M. and Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. Frontiers in Genetics, 5.

Blackwell, A. R., Dluzewska, J., Szymanska-Lejman, M., Desjardins, S., Tock, A. J., Kbiri, N., Lambing, C., Lawrence, E. J., Bieluszewski, T., Rowan, B., Higgins, J. D., Ziolkowski, P. A., and Henderson, I. R. (2020). Msh2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in arabidopsis. Embo Journal, 39(21).

Bradshaw, J. E. (2017). Plant breeding: past, present and future. Euphytica, 213(3).

Brown, T. (2002). Genomes 2nd ed.; Chapter 14: Mutation, Repair and Recombination, book section 14. Garland Science.

Brown, T. A. (2011). Introduction to Genetics: A Molecular Approach. Taylor & Francis Ltd.

Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K. K., Kumar, A., Pradhan, A. K., Kumar, U., Kumar, R. R., Juliana, P., Mishra, D. C., and Kumar, S. (2022). Genomic selection: A tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. Frontiers in Genetics, 13.

Burgess, A. J., Masclaux-Daubresse, C., Strittmatter, G., Weber, A. P. M., Taylor, S. H., Harbinson, J., Yin, X. Y., Long, S., Paul, M. J., Westhoff, P., Loreto, F., Ceriotti, A., Saltenis, V. L. R., Pribil, M., Nacry, P., Scharff, L. B., Jensen, P. E., Muller, B., Cohan, J. P., Foulkes, J., Rogowsky, P., Debaeke, P., Meyer, C., Nelissen, H., Inzé, D., Lankhorst, R. K., Parry, M. A. J., Murchie, E. H., and Baekelandt, A. (2023). Improving crop yield potential: Underlying biological processes and future prospects. Food and Energy Security, 12(1).

Burney, J. A., Davis, S. J., and Lobell, D. B. (2010). Greenhouse gas mitigation by agricultural intensification. Proceedings of the National Academy of Sciences of the United States of America, 107(26):12052–12057.

Casale, F., Van Inghelandt, D., Weisweiler, M., Li, J. Q., and Stich, B. (2022). Genomic prediction of the recombination rate variation in barley - a route to highly recombinogenic genotypes. Plant Biotechnology Journal, 20(4):676–690.

Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics, 8(10):762–775.

Choi, K. and Henderson, I. R. (2015). Meiotic recombination hotspots – a comparative view. The Plant Journal, 83(1):52–61.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C. H., McVean, G., and Henderson, I. R. (2013). Arabidopsis meiotic crossover hot spots overlap with h2a.z nucleosomes at gene promoters. Nature Genetics, 45(11):1327–1336.

Clark, S. A. and Van Der Werf, J. (2013). Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values, pages 321–330. Humana Press.

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. Biology, 1(2):439–459.

Crick, F. H. C., Griffith, J. S., and Orgel, L. E. (1957). Codes without commas. Proceedings of the National Academy of Sciences of the United States of America, 43(5):416–421.

Dalrymple, D. G. (1980). Development and spread of semi-dwarf varieties of wheat and rice in the united states: An international perspective. Report, United States Department of Agriculture, Economic Research Service. Agricultural Economic Reports.

Daniel, A., Vincent, G., Julie, R., Bastien, M., Antonio, C. A. M.-F., Romane, B., David, H., and Bart, D. (2019). Brb-seq: ultra-affordable high-throughput transcriptomics enabled by bulk rna barcoding and sequencing. Genome Biology, 20(1):71.

Danovi, S. (2023). Mutation rates across species. Nature Genetics, 55(4):524–524.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection. John Murray, London.

Darwin, C. (1900). The effects of cross and self fertilisation in the vegetable kingdom. J. Murray, London, 2d. 5th impression edition. 1st edition, 1876.

Dekkers, J. C. and Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet, 3(1):22–32.

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., and Batley, J. (2020). Characterization of disease resistance genes in the pangenome reveals significant structural variation. Plant Biotechnology Journal, 18(4):969–982.

Dreissig, S., Mascher, M., and Heckmann, S. (2019). Variation in recombination rate is shaped by domestication and environmental conditions in barley. Molecular Biology and Evolution, 36(9):2029–2039.

Dreissig, S., Maurer, A., Sharma, R., Milne, L., Flavell, A. J., Schmutzer, T., and Pillen, K. (2020). Natural variation in meiotic recombination rate shapes introgression patterns in intraspecific hybrids between wild and domesticated barley. New Phytologist, 228(6):1852–1863.

Duvick, D. N. and Cassman, K. G. (1999). Post–green revolution trends in yield potential of temperate maize in the north-central united states. Crop Science, 39(6):1622–1630.

Elaine, R. M. (2011). A decade's perspective on dna sequencing technology. Nature, 470(7333):198–203.

Elisabetta, M., Atefeh, L., Catia, M., Christoph, Z., Davis, J. M., Adrián, A.-V., Eduard, B., Sagar, Dominic, G., Julia, K. L., Stéphane, C. B., Chad, S., Aik, O., Robert, C. J., Kelly, K., Chris, B., Yasha, T., Yohei, S., Kaori, T., Tetsutaro, H., Caroline, B., Cornelius, F., Sascha, S., Timo, T., Christian, C., Xian, A., Lan, T. N., Aviv, R., Joshua, Z. L., Swati, P., Aleksandar, J., Lucas, E. W., Johannes, W. B., Wolfgang, E., Marta, G., Rickard, S., Itoshi, N., Ivo, G., Oliver, S., and Holger, H. (2020). Benchmarking single-cell rna-sequencing protocols for cell atlas projects. Nature Biotechnology, 38(6):747–755.

Evan, Z. M., Anindita, B., Rahul, S., James, N., Karthik, S., Melissa, G., Itay, T., Allison, R. B., Nolan, K., Emily, M. M., John, J. T., David, A. W., Joshua, R. S., Alex, K. S., Aviv, R., and Steven, A. M. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5):1202–1214.

FAO (2004). Agricultural biotechnology, meeting the needs of the poor?

Fernandes, J. B., Naish, M., Lian, Q. C., Burns, R., Tock, A. J., Rabanal, F. A., Wlodzimierz, P., Habring, A., Nicholas, R. E., Weigel, D., Mercier, R., and Henderson, I. R. (2024). Structural variation and dna methylation shape the centromere-proximal meiotic crossover landscape in arabidopsis. Genome Biology, 25(1).

Fernando, R. L. and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution, 21(4):467–477.

Friedt, W. and Ordon, F. (2022). Mendel's laws and their impact on plant breeding. Themenheft: Von Mendel zur Genomeditierung, 74(11-12):223–232.

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., and Alexandrov, N. (2019). Structural variants in 3000 rice genomes. Genome Research, 29(5):870–880.

Gao, L., Gonda, I., Sun, H. H., Ma, Q. Y., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y. M., van der Knaap, E., Huang, S. W., Klee, H. J., Giovannoni, J. J., and Fei, Z. J. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature Genetics, 51(6):1044–+.

Gayon, J. (2016). From mendel to epigenetics: History of genetics. Comptes Rendus. Biologies, 339(7-8):225–230.

Gion, J. M., Hudson, C. J., Lesur, I., Vaillancourt, R. E., Potts, B. M., and Freeman, J. S. (2016). Genome-wide variation in recombination rate in. Bmc Genomics, 17.

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C., and Mézard, C. (2011). Genome-wide crossover distribution in arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. PLoS Genetics, 7(11):e1002354.

Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002). Finding genes that underlie complex traits. Science, 298(5602):2345–9.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., and Edwards, D. (2016). The pangenome of an agronomically important crop plant brassica oleracea. Nature Communications, 7(1):13390.

Griffiths, A., Doebley, J., Peichel, C., and Wassarman, D. (2020). Introduction to Genetic Analysis. W. H. Freeman, 12th edition edition.

Guo, Z. G., Magwire, M. M., Basten, C. J., Xu, Z. Y., and Wang, D. L. (2016). Evaluation of the utility of gene expression and metabolic information for ge-

nomic prediction in maize. Theoretical and Applied Genetics, 129(12):2413–2427.

Harfouche, A. L., Jacobson, D. A., Kainer, D., Romero, J. C., Harfouche, A. H., Mugnozza, G. S., Moshelion, M., Tuskan, G. A., Keurentjes, J. J. B., and Altman, A. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends in Biotechnology, 37(11):1217–1235.

Hargrove, T. R. and Cabanilla, V. L. (1979). The impact of semidwarf varieties on asian rice-breeding programs. BioScience, 29(12):731–735.

Hassan, S., Surakka, I., Taskinen, M. R., Salomaa, V., Palotie, A., Wessman, M., Tukiainen, T., Pirinen, M., Palta, P., and Ripatti, S. (2021). High-resolution population-specific recombination rates and their effect on phasing and genotype imputation. European Journal of Human Genetics, 29(4):615–624.

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. Crop Science, 49(1):1–12.

Henderson, C. R. (1984). Applications of linear models in animal breeding. University of Guelph, Guelph.

Henderson, I. R. (2012). Control of meiotic recombination frequency in plant genomes. Current Opinion in Plant Biology, 15(5):556–561.

Higgins, J. D., Perry, R. M., Barakate, A., Ramsay, L., Waugh, R., Halpin, C., Armstrong, S. J., and Franklin, F. C. H. (2012). Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in barley. The Plant Cell, 24(10):4096–4109.

Hisano, H., Sakamoto, K., Takagi, H., Terauchi, R., and Sato, K. (2017). Exome qtl-seq maps monogenic locus and qtls in barley. BMC Genomics, 18(1):125.

Hsu, Y.-M., Falque, M., and Martin, O. C. (2022). Quantitative modelling of fine-scale variations in the arabidopsis thaliana crossover landscape. Quantitative Plant Biology, 3.

Hu, H., Campbell, M. T., Yeats, T. H., Zheng, X., Runcie, D. E., Covarrubias-Pazaran, G., Broeckling, C., Yao, L., Caffe-Treml, M., Gutierrez, L. A., Smith, K. P., Tanaka, J., Hoekenga, O. A., Sorrells, M. E., Gore, M. A., and Jannink, J. L. (2021). Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. Theor Appl Genet, 134(12):4043–4054.

Hu, X., Carver, B. F., El-Kassaby, Y. A., Zhu, L., and Chen, C. (2023). Weighted kernels improve multi-environment genomic prediction. Heredity (Edinb), 130(2):82–91.

Irene Gallego, R., Athma, A. P., Jenny, T., and Yoav, G. (2014). Rna-seq: impact of rna degradation on transcript quantification. BMC Biology, 12(1):42.

Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M. T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Bekele, W., Brar, G. S., Braune, K., Bunk, B., Chalmers, K. J., Chapman, B., Jørgensen, M. E., Feng, J.-W., Feser, M., Fiebig, A., Gundlach, H., Guo, W., Haberer, G., Hansson, M., Himmelbach, A., Hoffie, I., Hoffie, R. E., Hu, H., Isobe, S., König, P., Kale, S. M., Kamal, N., Keeble-Gagnère, G., Keller, B., Knauft, M., Koppolu, R., Krattinger, S. G., Kumlehn, J., Langridge, P., Li, C., Marone, M. P., Maurer, A., Mayer, K. F. X., Melzer, M., Muehlbauer, G. J., Murozuka, E., Padmarasu, S., Perovic, D., Pillen, K., Pin, P. A., Pozniak, C. J., Ramsay, L., Pedas, P. R., Rutten, T., Sakuma, S., Sato, K., Schüler, D., Schmutzer, T., Scholz, U., Schreiber, M., Shirasawa, K., Simpson, C., Skadhauge, B., Spannagl, M., Steffenson, B. J., Thomsen, H. C., Tibbits, J. F., Nielsen, M. T. S., Trautewig, C., Vequaud, D., Voss, C., Wang, P., Waugh, R., Westcott, S., Rasmussen, M. W., Zhang, R., Zhang, X.-Q., Wicker, T., Dockter, C., Mascher, M., and Stein, N. (2024). Structural variation in the pangenome of wild and domesticated barley. Nature, 636(8043):654–662.

Jeffery, M. V., Kathryn, S., Mark, K. D., Elke, A. J., Andrew, P. S., and Jason, G. (2019). Ffpecap-seq: A method for sequencing capped rnas in formalin-fixed paraffin-embedded samples. Genome Research, 29(11):1826–1835.

Jesse, D. and Matthias, M. (2012). Length and gc-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern dna sequencing libraries. BioTechniques, 52(2).

Jia, Y. and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics, 192(4):1513–22.

Johannes, W. B., Christoph, Z., Aleksandar, J., Lucas, E. W., Beate, V., Swati, P., Johanna, G., Ines, H., and Wolfgang, E. (2018). Sensitive and powerful single-cell rna sequencing using mcscrb-seq. Nature Communications, 9(1).

Jordan, K. W., Wang, S. C., He, F., Chao, S. A. M., Lun, Y. N., Paux, E., Sourdille, P., Sherman, J., Akhunova, A., Blake, N. K., Pumphrey, M. O., Glover, K., Dubcovsky, J., Talbert, L., and Akhunov, E. D. (2018). The genetic

architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. Plant Journal, 95(6):1039–1054.

Joseph, W. F., Chunfang, Z., Philippe, J., Shirley, X. Z., Peipei, L., Michael, J. M., and Robert, B. W. (2019). Gene expression profiling of single cells from archival tissue with laser-capture microdissection and smart-3seq. Genome Research, 29(11):1816–1825.

Khush, G. S. (2001). Green revolution: the way forward. Nature Reviews Genetics, 2(10):815–822.

Kono, T. J. Y., Fu, F., Mohammadi, M., Hoffman, P. J., Liu, C., Stupar, R. M., Smith, K. P., Tiffin, P., Fay, J. C., and Morrell, P. L. (2016). The role of deleterious substitutions in crop genomes. Molecular Biology and Evolution, 33(9):2307–2317.

Kris, A. W. (2021). The cost of sequencing a human genome.

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124(3):743–56.

Lee, S.-I. and Kim, N.-S. (2014). Transposable elements and genome size variations in plants. Genomics & Informatics, 12(3):87.

Lei, S. A. (2010). Benefits and costs of vegetative and sexual reproduction in perennial plants: A review of literature. Journal of the Arizona-Nevada Academy of Science, 42(1):9–14.

Lenaerts, B., Collard, B. C. Y., and Demont, M. (2019). Review: Improving global food security through accelerated plant breeding. Plant Science, 287.

Li, D., Xu, Z., Gu, R., Wang, P., Lyle, D., Xu, J., Zhang, H., and Wang, G. (2019). Enhancing genomic selection by fitting large-effect snps as fixed effects and a genotype-by-environment effect using a maize bc1f3:4 population. PLoS One, 14(10):e0223898.

Li, X., Li, L., and Yan, J. B. (2015). Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. Nature Communications, 6.

Li, Y., Xiao, J., Wu, J., Duan, J., Liu, Y., Ye, X., Zhang, X., Guo, X., Gu, Y., Zhang, L., Jia, J., and Kong, X. (2012). A tandem segmental duplication (tsd) in green revolution gene rht-d1b region underlies plant height variation. New Phytologist, 196(1):282–291.

Louise, A., Yong, G., and Michael, A. Q. (2016). Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital pcr assays - a systematic comparison of dna library preparation kits for illumina sequencing. BMC Genomics, 17(1):458.

Lu, P. L., Han, X. W., Qi, J., Yang, J. G., Wijeratne, A. J., Li, T., and Ma, H. (2012). Analysis of arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing landsberg erecta and all four products of a single meiosis. Genome Research, 22(3):508–518.

Luisa, F. P., Serge, P., and Julien, F. A. (2019). Tm3'seq: A tagmentation-mediated 3' sequencing approach for improving scalability of rnaseq experiments. G3: Genes, Genomes, Genetics, page g3.400821.2019.

Lyra, D. H., Mendonça, L. D., Galli, G., Alves, F. C., Granato, I. S. C., and Fritsche-Neto, R. (2017). Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. Molecular Breeding, 37(6).

Mackay, T. F. (2001). The genetic architecture of quantitative traits. Annu Rev Genet, 35:303–39.

Madeline, Y. M., Lillian, M. K., Eric, D. C., Matt, S. Z., Joseph, L. D., and Torsten, T. (2019). Miniaturization and optimization of 384-well compatible rna sequencing library preparation. PLoS ONE, 14(1):e0206194.

Mahmoud, M., Gobet, N., Cruz-Davalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. Genome Biol, 20(1):246.

Marand, A. P., Zhao, H. N., Zhang, W. L., Zeng, Z. X., Fang, C., and Jiang, J. M. (2019). Historical meiotic crossover hotspots fueled patterns of evolutionary divergence in rice. Plant Cell, 31(3):645–662.

Maron, L. G., Guimaraes, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., Buckler, E. S., Coluccio, A. E., Danilova, T. V., Kudrna, D., Magalhaes, J. V., Pineros, M. A., Schatz, M. C., Wing, R. A., and Kochian, L. V. (2013). Aluminum tolerance in maize is associated with higher mate1 gene copy number. Proc Natl Acad Sci U S A, 110(13):5241–6.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P. E., Russell, J., Bayer, M., Ramsay, L., Liu, H., Haberer, G., Zhang, X. Q., Zhang, Q., Barrero, R. A., Li, L., Taudien, S., Groth, M., Felder, M., Hastie, A., Simkova, H., Stankova, H., Vrana, J.,

Chan, S., Munoz-Amatriain, M., Ounit, R., Wanamaker, S., Bolser, D., Colmsee, C., Schmutzer, T., Aliyeva-Schnorr, L., Grasso, S., Tanskanen, J., Chailyan, A., Sampath, D., Heavens, D., Clissold, L., Cao, S., Chapman, B., Dai, F., Han, Y., Li, H., Li, X., Lin, C., McCooke, J. K., Tan, C., Wang, P., Wang, S., Yin, S., Zhou, G., Poland, J. A., Bellgard, M. I., Borisjuk, L., Houben, A., Dolezel, J., Ayling, S., Lonardi, S., Kersey, P., Langridge, P., Muehlbauer, G. J., Clark, M. D., Caccamo, M., Schulman, A. H., Mayer, K. F. X., Platzer, M., Close, T. J., Scholz, U., Hansson, M., Zhang, G., Braumann, I., Spannagl, M., Li, C., Waugh, R., and Stein, N. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature, 544(7651):427–433.

Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A., Matsumoto, T., Sato, K., Schulman, A., Muehlbauer, G. J., Stein, N., Ariyadasa, R., Schulte, D., Poursarebani, N., Zhou, R., Steuernagel, B., Mascher, M., Scholz, U., Shi, B., Langridge, P., Madishetty, K., Svensson, J. T., Bhat, P., Moscou, M., Resnik, J., Close, T. J., Muehlbauer, G. J., Hedley, P., Liu, H., Morris, J., Waugh, R., Frenkel, Z., Korol, A., Bergès, H., Graner, A., Stein, N., Steuernagel, B., Scholz, U., Taudien, S., Felder, M., Groth, M., Platzer, M., Stein, N., Steuernagel, B., Scholz, U., Himmelbach, A., Taudien, S., Felder, M., Platzer, M., Lonardi, S., Duma, D., Alpert, M., Cordero, F., Beccuti, M., Ciardo, G., Ma, Y., Wanamaker, S., Close, T. J., Stein, N., Cattonaro, F., Vendramin, V., Scalabrin, S., Radovic, S., Wing, R., Schulte, D., Steuernagel, B., Morgante, M., Stein, N., Waugh, R., Nussbaumer, T., Gundlach, H., Martis, M., Ariyadasa, R., Poursarebani, N., Steuernagel, B., Scholz, U., Wise, R. P., Poland, J., Stein, N., Mayer, K. F. X., Spannagl, M., Pfeifer, M., Gundlach, H., Mayer, K. F. X., Gundlach, H., Moisy, C., Tanskanen, J., Scalabrin, S., Zuccolo, A., Vendramin, V., Morgante, M., Mayer, K. F. X., Schulman, A., Pfeifer, M., Spannagl, M., Hedley, P., Morris, J., Russell, J., Druka, A., Marshall, D., et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature, 491(7426):711–716.

McColgan, P. and Tabrizi, S. J. (2018). Huntington's disease: a clinical review. Eur J Neurol, 25(1):24–34.

McCombie, W. R., John, D. M., and Elaine, R. M. (2019). Next-generation sequencing technologies. Cold Spring Harbor Perspectives in Medicine, 9(11).

McMahill, M. S., Sham, C. W., and Bishop, D. K. (2007). Synthesis-dependent strand annealing in meiosis. PLoS Biol, 5(11):e299.

Mendel, G. and Bateson, W. (1925). Experiments in plant-hybridisation. Harvard

University Press, Cambridge, Mass. This translation of Mendel's Versuche über Pflanzen-Hybriden (Abh. Naturf. ver. Brünn, v. 4, 1866) by the Royal Horticultural Society, London, appeared in their Journal, v. 26, p. 1, 1901, from which.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., and Grelon, M. (2015). The molecular biology of meiosis in plants. Annual Review of Plant Biology, Vol 66, 66:297–327.

Merrick, L. F. and Carter, A. H. (2021). Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. Plant Genome, 14(3):e20158.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4):1819–1829.

Michael, G. R., Carsten, R., Maura, C., Andrew, H., Niall, J. L., Ryan, H., Chad, N., and David, B. J. (2013). Characterizing and measuring bias in sequence data. Genome Biology, 14(5):R51.

Michel, S., Wagner, C., Nosenko, T., Steiner, B., Samad-Zamini, M., Buerstmayr, M., Mayer, K., and Buerstmayr, H. (2021). Merging genomics and transcriptomics for predicting fusarium head blight resistance in wheat. Genes, 12(1).

Migicovsky, Z. and Kovalchuk, I. (2013). Changes to dna methylation and homologous recombination frequency in the progeny of stressed plants. Biochemistry and Cell Biology, 91(1):1–5.

Miles, C. and Wayne, M. (2008). Quantitative trait locus (qtl) analysis. Nature Education 1(1):208.

Mochida, K., Nishii, R., and Hirayama, T. (2020). Decoding plant-environment interactions that influence crop agronomic traits. Plant Cell Physiol, 61(8):1408–1418.

Montesinos-Lopez, O. A., Montesinos-Lopez, J. C., Salazar, E., Barron, J. A., Montesinos-Lopez, A., Buenrostro-Mariscal, R., and Crossa, J. (2021). Application of a poisson deep neural network model for the prediction of count data in genome-based prediction. Plant Genome, 14(3).

Morgan, T. H., Sturtevant, A. H., Muller, H. J., and Bridges, C. B. (1923). The mechanism of Mendelian heredity. New York : Henry Holt and Company.

Muller, H. J. (1927). Artificial transmutation of the gene. Science, 66(1699):84–87.

Muller, H. J. (1928). The measurement of gene mutation rate in drosophila, its high variability, and its dependence upon temperature. Genetics, 13(4):279–357.

Munoz-Amatriain, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K. F., Taudien, S., Platzer, M., Jeddeloh, J. A., Springer, N. M., Muehlbauer, G. J., and Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol, 14(6):R58.

Newton, A. C., Flavell, A. J., George, T. S., Leat, P., Mullholland, B., Ramsay, L., Revoredo-Giha, C., Russell, J., Steffenson, B. J., Swanston, J. S., Thomas, W. T. B., Waugh, R., White, P. J., and Bingham, I. J. (2011). Crops that feed the world 4. barley: a resilient crop? strengths and weaknesses in the context of food security. Food Security, 3(2):141–178.

Nishida, H., Yoshida, T., Kawakami, K., Fujita, M., Long, B., Akashi, Y., Laurie, D. A., and Kato, K. (2013). Structural variation in the 5' upstream region of photoperiod-insensitive alleles ppd-a1a and ppd-b1a identified in hexaploid wheat (triticum aestivum l.), and their effect on heading time. Molecular Breeding, 31(1):27–37.

Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N., and Tiffin, P. (2012). Fine-scale population recombination rates, hotspots, and correlates of recombination in the genome. Genome Biology and Evolution, 4(5):726–737.

Parks, M. M., Lawrence, C. E., and Raphael, B. J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. Genome Biology, 16(1).

Peck, J. R. (1994). A ruby in the rubbish - beneficial mutations, deleterious mutations and the evolution of sex. Genetics, 137(2):597–606.

Pingali, P. L. (2012). Green revolution: Impacts, limits, and the path ahead. Proceedings of the National Academy of Sciences of the United States of America, 109(31):12302–12308.

Pollard, T. D., Earnshaw, W. C., Lippincott-Schwartz, J., and Johnson, G. T., editors (2017). Chapter 45 - Meiosis, pages 779–795. Elsevier.

Pourkheirandish, M. and Komatsuda, T. (2007). The importance of barley genetics and domestication in a global perspective. Annals of Botany, 100(5):999–1008.

Purugganan, M. D. and Fuller, D. Q. (2009). The nature of selection during plant domestication. Nature, 457(7231):843–848.

Ritz, K. R., Noor, M. A. F., and Singh, N. D. (2017). Variation in recombination rate: Adaptive or not? Trends Genet, 33(5):364–374.

Rochette, N. C., Rivera-Colon, A. G., Walsh, J., Sanger, T. J., Campbell-Staton, S. C., and Catchen, J. M. (2023). On the causes, consequences, and avoidance of pcr duplicates: Towards a theory of library complexity. Mol Ecol Resour, 23(6):1299–1318.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C. H., Li, Y. X., and Buckler, E. S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences of the United States of America, 112(12):3823–3828.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R., and Weigel, D. (2019). An ultra high-density arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. Genetics, 213(3):771–787.

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., and Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3-Genes Genomes Genetics, 6(9):2799–2808.

Saintenac, C., Faure, S., Remay, A., Choulet, F., Ravel, C., Paux, E., Balfourier, F., Feuillet, C., and Sourdille, P. (2011). Variation in crossover rates across a 3-mb contig of bread wheat (triticum aestivum) reveals the presence of a meiotic recombination hotspot. Chromosoma, 120(2):185–198.

Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., and Weigel, D. (2012). The recombination landscape in f populations. Heredity, 108(4):447–455.

Samantha, N. M., Patrick, R. M., Joshua, A. R., Eric, J. D., and John, D. P. (2020). Impact of reducing dna input on next-generation sequencing library complexity and variant detection. The Journal of Molecular Diagnostics, 22(5):720–727.

Samuel, M., Austin, H., Alexei, S., Noelani, K., Vincent, L. B., and Stuart, S. L. (2020). High-throughput minitaturized rna-seq library preparation. Journal of Biomolecular Techniques, 31(4):151–156.

Sandhu, K., Patil, S. S., Pumphrey, M., and Carter, A. (2021). Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. Plant Genome, 14(3).

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., and Melchinger, A. E. (2018). Beyond genomic prediction: Combining different types of data can improve prediction of hybrid performance in maize. Genetics, 208(4):1373–1385.

Schuermann, D., Molinier, J., Fritsch, O., and Hohn, B. (2005). The dual nature of homologous recombination in plants. Trends in Genetics, 21(3):172–181.

Schule, B., McFarland, K. N., Lee, K., Tsai, Y. C., Nguyen, K. D., Sun, C., Liu, M., Byrne, C., Gopi, R., Huang, N., Langston, J. W., Clark, T., Gil, F. J. J., and Ashizawa, T. (2017). Parkinson's disease associated with pure atxn10 repeat expansion. NPJ Parkinsons Dis, 3:27.

Sergio, M.-C., Cuong, T., Soheila, V., Robert, M., Srimeenakshi, S., Jennifer, N. D., Heidi, C.-A., Joby, J., and Louise, C. L. (2016). Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. Journal of Laboratory Automation, 21(4):557–567.

Si, W. N., Yuan, Y., Huang, J., Zhang, X. H., Zhang, Y. C., Zhang, Y. D., Tian, D. C., Wang, C. L., Yang, Y. H., and Yang, S. H. (2015). Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice f plants. New Phytologist, 206(4):1491–1502.

Silva, O. B. and Grattapaglia, D. (2015). Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of. New Phytologist, 208(3):830–845.

Soumillon, M., Cacchiarelli, D., Semrau, S., Oudenaarden, A. v., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell rna-seq. bioRxiv, page 003236.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H., Konkel, M. K., Malhotra, A., Stutz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan,

X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E. W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., Genomes Project, C., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571):75–81.

Sun, Y. J., Ambrose, J. H., Haughey, B. S., Webster, T. D., Pierrie, S. N., Muñoz, D. F., Wellman, E. C., Cherian, S., Lewis, S. M., Berchowitz, L. E., and Copenhaver, G. P. (2012). Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in. Plos Genetics, 8(10).

Sutton, T., Baumann, U., Hayes, J., Collins, N. C., Shi, B. J., Schnurbusch, T., Hay, A., Mayo, G., Pallotta, M., Tester, M., and Langridge, P. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science, 318(5855):1446–9.

Tahakik, R. R., Deshmukh, A. G., Moharil, M. P., Jadhav, P. V., Kogade, V. T., More, K. D., and Shinde, V. P. (2024). Transitioning from the green revolution to the gene revolution: strengthening nutritional security using climate resilient traditional crops. Bulletin of the National Research Centre, 48(1).

Tamar, H., Naftalie, S., Gal, A., Agnes, K., Yaron de, L., Leon, A., Dave, G., Shuqiang, L., Kenneth, J. L., Orit, R.-R., Yuval, D., Aviv, R., and Itai, Y. (2016). Cel-seq2: Sensitive highly-multiplexed single-cell rna-seq. Genome Biology, 17(1):77.

Tilman, D., Balzer, C., Hill, J., and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences of the United States of America, 108(50):20260–20264.

Tsai, H. Y., Cericola, F., Edriss, V., Andersen, J. R., Orabiid, J., Jensen, J. D., Jahoor, A., Janss, L., and Jensen, J. (2020). Use of multiple traits genomic prediction, genotype by environment interactions and spatial effect to improve prediction accuracy in yield data. Plos One, 15(5).

Vicient, C. M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharav, A., Nevo, E., and Schulman, A. H. (1999). Retrotransposon bare-1 and its role in genome evolution in the genus hordeum. The Plant Cell, 11(9):1769–1784.

Visioni, A., Basile, B., Amri, A., Sanchez-Garcia, M., and Corrado, G. (2023). Advancing the conservation and utilization of barley genetic resources: Insights into germplasm management and breeding for sustainable agriculture. Plants, 12(18):3186.

von Diezmann, L. and Rog, O. (2021). Let's get physical - mechanisms of crossover interference. J Cell Sci, 134(10).

Walkowiak, S., Gao, L. L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., Ban, T., Venturini, L., Bevan, M., Clavijo, B., Koo, D. H., Ens, J., Wiebe, K., N'Diaye, A., Fritz, A. K., Gutwin, C., Fiebig, A., Fosker, C., Fu, B. X., Accinelli, G. G., Gardner, K. A., Fradgley, N., Gutierrez-Gonzalez, J., Halstead-Nussloch, G., Hatakeyama, M., Koh, C. S., Deek, J., Costamagna, A. C., Fobert, P., Heavens, D., Kanamori, H., Kawaura, K., Kobayashi, F., Krasileva, K., Kuo, T., McKenzie, N., Murata, K., Nabeka, Y., Paape, T., Padmarasu, S., Percival-Alwyn, L., Kagale, S., Scholz, U., Sese, J., Juliana, P., Singh, R., Shimizu-Inatsugi, R., Swarbreck, D., Cockram, J., Budak, H., Tameshige, T., Tanaka, T., Tsuji, H., Wright, J., Wu, J. Z., Steuernagel, B., Small, I., Cloutier, S., Keeble-Gagnère, G., Muehlbauer, G., Tibbets, J., Nasuda, S., Melonek, J., Hucl, P. J., Sharpe, A. G., Clark, M., Legg, E., Bharti, A., Langridge, P., Hall, A., Uauy, C., Mascher, M., Krattinger, S. G., Handa, H., Shimizu, K. K., Distelfeld, A., Chalmers, K., Keller, B., Mayer, K. F. X., Poland, J., Stein, N., McCartney, C. A., Spannagl, M., Wicker, T., and Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837).

Wang, H. (2023). Case studies of breeding strategies in major plant species.

Wang, J., Street, N. R., Scofield, D. G., and Ingvarsson, P. K. (2016). Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related. Genetics, 202(3):1185–+.

Washburn, J. D., Burch, M. B., and Franco, J. A. V. (2020). Predictive breeding for maize: Making use of molecular phenotypes, machine learning, and physiological crop models. Crop Science, 60(2):622–638.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. Nature, 171(4356):737–738.

Wei, Z., Lisa, M. C., and Hongyu, Z. (2011). Bias detection and correction in rna-sequencing data. BMC Bioinformatics, 12(1):290.

Weisweiler, M., de Montaigu, A., Ries, D., Pfeifer, M., and Stich, B. (2019). Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mrna sequencing and their power to predict phenotypic traits. Bmc Genomics, 20(1).

Westhues, M., Heuer, C., Thaller, G., Fernando, R., and Melchinger, A. E. (2019). Efficient genetic value prediction using incomplete omics data. Theoretical and Applied Genetics, 132(4):1211–1222.

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., Stitt, M., Nikoloski, Z., Willmitzer, L., Schön, C. C., Scholten, S., and Melchinger, A. E. (2017). Omics-based hybrid prediction in maize. Theoretical and Applied Genetics, 130(9):1927–1939.

Wicker, T., Schulman, A. H., Tanskanen, J., Spannagl, M., Twardziok, S., Mascher, M., Springer, N. M., Li, Q., Waugh, R., Li, C., Zhang, G., Stein, N., Mayer, K. F. X., and Gundlach, H. (2017). The repetitive landscape of the 5100 mbp barley genome. Mob DNA, 8:22.

Wieczorek, A. M. & Wright, M. G. (2012). History of agricultural biotechnology: How crop development has evolved. Nature Education Knowledge 3(10):9.

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L., Huettel, B., de Jong, H., Ossowski, S., Weigel, D., Koornneef, M., Keurentjes, J. J., and Schneeberger, K. (2013). The genomic landscape of meiotic crossovers and gene conversions in arabidopsis thaliana. Elife, 2:e01426.

Wu, P. Y., Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., and Van Inghelandt, D. (2022). Improvement of prediction ability by integrating multi-omic datasets in barley. Bmc Genomics, 23(1).

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., Dong, Y., Gutenkunst, R. N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J., and Wang, W. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology, 30(1):105–111.

Yang, S. H., Yuan, Y., Wang, L., Li, J., Wang, W., Liu, H. X., Chen, J. Q., Hurst, L. D., and Tian, D. C. (2012). Great majority of recombination events in are

gene conversion events. Proceedings of the National Academy of Sciences of the United States of America, 109(51):20992–20997.

Yuan, Y. X., Bayer, P. E., Batley, J., and Edwards, D. (2021). Current status of structural variation studies in plants. Plant Biotechnology Journal, 19(11):2153–2163.

Zabel, F., Delzeit, R., Schneider, J. M., Seppelt, R., Mauser, W., and Václavík, T. (2019). Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity. Nature Communications, 10.

Zhang, M., Liu, Y.-H., Xu, W., Smith, C. W., Murray, S. C., and Zhang, H.-B. (2020). Analysis of the genes controlling three quantitative traits in three diverse plant species reveals the molecular basis of quantitative traits. Scientific Reports, 10(1).

Zhao, Y., Mette, M. F., Gowda, M., Longin, C. F. H., and Reif, J. C. (2014). Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity, 112(6):638–645.

3 Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation

This manuscript was published in Theoretical and Applied Genetics in August 2022. Supplementary material is available online.

Authors:
Marius Weisweiler, **Christopher Arlt**, Po-Ya Wu, Delphine Van Inghelandt, Thomas Hartwig, and Benjamin Stich.

Own contribution: Second author. I performed the SV validation in the lab and contributed to the manuscript in the corresponding sections.

Theoretical and Applied Genetics (2022) 135:3511–3529
https://doi.org/10.1007/s00122-022-04197-7

ORIGINAL ARTICLE

Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation

Marius Weisweiler[1] · Christopher Arlt[1] · Po-Ya Wu[1] · Delphine Van Inghelandt[1] · Thomas Hartwig[2] · Benjamin Stich[1,3]

Received: 11 May 2022 / Accepted: 3 August 2022 / Published online: 27 August 2022

Abstract

Key message **Structural variants (SV) of 23 barley inbreds, detected by the best combination of SV callers based on short-read sequencing, were associated with genome-wide and gene-specific gene expression and, thus, were evaluated to predict agronomic traits.**

Abstract In human genetics, several studies have shown that phenotypic variation is more likely to be caused by structural variants (SV) than by single nucleotide variants. However, accurate while cost-efficient discovery of SV in complex genomes remains challenging. The objectives of our study were to (i) facilitate SV discovery studies by benchmarking SV callers and their combinations with respect to their sensitivity and precision to detect SV in the barley genome, (ii) characterize the occurrence and distribution of SV clusters in the genomes of 23 barley inbreds that are the parents of a unique resource for mapping quantitative traits, the double round robin population, (iii) quantify the association of SV clusters with transcript abundance, and (iv) evaluate the use of SV clusters for the prediction of phenotypic traits. In our computer simulations based on a sequencing coverage of 25x, a sensitivity > 70% and precision > 95% was observed for all combinations of SV types and SV length categories if the best combination of SV callers was used. We observed a significant ($P < 0.05$) association of gene-associated SV clusters with global gene-specific gene expression. Furthermore, about 9% of all SV clusters that were within 5 kb of a gene were significantly ($P < 0.05$) associated with the gene expression of the corresponding gene. The prediction ability of SV clusters was higher compared to that of single-nucleotide polymorphisms from an array across the seven studied phenotypic traits. These findings suggest the usefulness of exploiting SV information when fine mapping and cloning the causal genes underlying quantitative traits as well as the high potential of using SV clusters for the prediction of phenotypes in diverse germplasm sets.

Introduction

Researchers began to study genomic rearrangements and structural variants (SV) about 60 years ago. These studies investigated somatic chromosomes, biopsies, and cell cultures from lymphomas to understand the role of abnormal chromosome numbers as well as SV for the development of cancer (Jacobs and Strong 1959; Nowell and Hungerford 1960; Manolov and Manolov 1972; Craig-Holmes et al. 1973; Mitelman et al. 1979).

The development of sequencing by synthesis pioneered by Frederick Sanger (Sanger et al. 1977) enabled in the following years the first sequenced genomes of prokaryotes (e.g., *Escherichia coli*) and eukaryotes (e.g., yeast) (Goffeau et al. 1996; Blattner et al. 1997). Next milestones of sequencing by synthesis were the sequenced genomes of

Communicated by Takao Komatsuda.

Christopher Arlt and Po-Ya Wu authors contributed equally.

✉ Benjamin Stich
benjamin.stich@hhu.de

[1] Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany

[2] Institute for Molecular Physiology, Universitätsstraße 1, 40225 Düsseldorf, Germany

[3] Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany

Arabidopsis thaliana as first plant species (The Arabidopsis Genome Iniative 2000) and of human (Craig Venter et al. 2001). Due to the development of next-generation sequencing (NGS) platforms such as 454 and Illumina, studies aiming for genome-wide variant detection in 100s or 1000s of samples as in the 1000 genome project (Altshuler et al. 2012) became possible.

Three different approaches have been proposed to detect SV based on NGS data: assembling, long-read sequencing, and short-read sequencing (Mahmoud et al. 2019). For crop and especially for cereal species, the assembly approach is a tough challenge because of the large genome size and the high proportion of repetitive elements in the genomes (Neale et al. 2014; Mascher et al. 2017). Long-read mapping requires Pacific Biosciences or Nanopore sequencing data which results in high costs if many accessions should be sequenced and, thus, is not affordable for many research groups. In contrast, short-read sequencing is well-established for SV detection in the human genome (Chaisson et al. 2019; Ebert et al. 2021). Various software tools have been developed to detect SV from short-read sequencing data and were benchmarked based on human genomes (Cameron et al. 2019; Kosugi et al. 2019).

More recently there is also an increased interest in using such approaches for SV detection in plant genomes (Fuentes et al. 2019; Zhou et al. 2019; Guan et al. 2021). Fuentes et al. (2019) evaluated several SV callers to detect SV in the rice genome. However, no study evaluated the performance of SV callers for transposon-rich complex cereal genomes.

Several studies have examined the distribution and frequency of SV in the genomes of rice and maize (Wang et al. 2018; Yang et al. 2019; Kou et al. 2020). Despite the importance of cereals for human nutrition, only Jayakodi et al. (2020) performed a genome-wide study on SV in barley, with a focus on large SV in 20 barley accessions.

In humans, SV have been described to have an up to ~ 50fold stronger influence on gene expression than single nucleotide variants (SNV) (Chiang et al. 2017). SV also have been associated with changes in transcript abundance in plants such as in cucumber (Zhang et al. 2015), maize (Yang et al. 2019), tomato (Alonge et al. 2020), and soybean (Liu et al. 2020a). However, the role and frequency of SV in gene regulatory mechanisms in small grain cereals is widely unexplored.

In humans, several studies have shown that phenotypic variation is more likely to be caused by SV than by SNV (Alkan et al. 2011; Baker 2012; Sudmant et al. 2015; Schüle et al. 2017; McColgan and Tabrizi 2018). In plants, individual SV have been associated with traits such as aluminum tolerance in maize (Maron et al. 2013), disease resistance and domestication in rice (Xu et al. 2012), or plant height (Li et al. 2012) and heading date (Nishida et al. 2013) in wheat. In barley, individual SV have been associated with traits such as Boron toxicity tolerance (Sutton et al. 2007) and disease resistance (Muñoz-Amatriaín et al. 2013). In grapevine and rice, it has been shown that SV have a low variant frequency due to purifying selection (Zhou et al. 2019; Kou et al. 2020). However, few studies have examined the ability to predict quantitatively inherited phenotypic traits using SV in comparison to SNV.

The objectives of our study were to (i) facilitate SV discovery studies by benchmarking SV callers and their combinations with respect to their sensitivity and precision to detect SV in the barley genome, (ii) characterize the occurrence and distribution of SV clusters in the genomes of 23 barley inbreds that are the parents of a unique resource for mapping quantitative traits, the double round robin population (Casale et al. 2022), (iii) quantify the association of SV clusters with transcript abundance, and (iv) evaluate the use of SV clusters for the prediction of phenotypic traits.

Methods

Benchmarking of variant callers for detecting SV and INDELs in the barley genome

Computer simulations

We used Mutation-Simulator (version 2.0.3) (Kühl et al. 2021) to simulate INDELs, deletions, duplications, inversions, insertions, and translocations in the first chromosome of the Morex reference sequence v2 (Monat et al. 2019) as this was the genome sequence available when our study was performed. Furthermore, it is not expected that the reference version impacts the results of the simulations. In accordance with Fuentes et al. 2019, we considered five SV length categories for each of the above mentioned SV types (except translocations) (A: 50–300 bp; B: 0.3–5 kb; C: 5–50 kb; D: 50–250 kb; E: 0.25–1 Mb) plus INDELs (2-49bp). Translocations were simulated for 50 bp–1 Mb (ABCDE). We simulated SV with a mutation rate of $1.9x10^{-6}$ for the SV length categories A-C and INDELs, whereas mutation rates of $3.8x10^{-6}$ and $1.9x10^{-7}$ were assumed for SV length categories D and E, respectively. For each type of SV, we used BBMap's randomreads.sh (BBMap - Bushnell B. - http://sourceforge.net/projects/bbmap/) to simulate 2x150 bp Illumina reads with a sequencing coverage of 1.5x, 3x, 6x, 12.5x, 25x, and 65x as well as LRSim (version 1.0) (Luo et al. 2017) to simulate linked-reads with a sequencing coverage of 14x and 25x. Illumina- and linked-reads were simulated with a minimum, average, and maximum base quality of 25, 35, and 40, respectively.

Theoretical and Applied Genetics (2022) 135:3511–3529

SV detection

The simulated Illumina reads were mapped to the first chromosome of the Morex reference sequence v2 using BWA-MEM (version 0.7.15) whereas LongRanger align (version 2.2.2) was used for the simulated linked-reads. The SV callers Pindel (version 0.2.5b9) (Ye et al. 2009), Delly (version 0.8.1) (Rausch et al. 2012), GRIDSS (version 2.8.3) (Cameron et al. 2017), Manta (version 1.6.0) (Chen et al. 2016), Lumpy (smoove version 0.2.5) (Layer et al. 2014), and NGSEP (version 3.3.2) (Duitama et al. 2014) were used to identify SV based on the mapped reads. GATK's HaplotypeCaller (4.1.6.0) (Poplin et al. 2017), Pindel, and GRIDSS were used to detect INDELs. The workflow was implemented in Snakemake (version 5.10.0) (Köster et al. 2021). A SV call was only kept if it passed the built-in filter of the corresponding SV caller. For INDELs and all SV types and length categories, only homozygous, alternative variant calls were considered. Deletions annotated as "replacement" (RPL) by Pindel were removed. We calculated the sensitivity (1), precision (2), and the F1-score (3) as

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$F1\text{ - core} = 2 * (\text{Precision*Sensitivity}/\text{Precision} + \text{Sensitivity}) \quad (3)$$

for all combinations of SV types*SV callers, where TP was the number of true positives, FP the number of false positives, and FN the number of false negatives. For INDELs, a TP INDEL had break points that did differ ≤ 2 bp from those of the simulated INDEL and the length did differ by ≤ 5bp. For SV length category A, a TP SV had break points that did differ ≤ 10 bp from those of the simulated SV and the SV length did differ by ≤ 20 bp. For the other SV length categories, a TP SV had break points and length differences compared to the simulated SV of ≤ 50 bp. For insertions where no SV length was detected, the start of a TP insertion had a break point that did differ ≤ 10 bp from this of the simulated insertion. For translocations, a TP translocation had break points that did differ ≤ 50 bp from those of the simulated translocation.

We also evaluated combinations of SV callers for their precision and sensitivity to detect SV. The following procedure was used to decide for the combinations that were examined: First, for those SV callers, which have shown a precision ≥ 95% for all SV length categories for a particular SV type, SV calls were combined via logical or ("|"). Second, for those SV callers with a precision ≤ 95% in at least one SV length category, SV calls were combined with a logical and ("&"). If the precision of the combination of the second step increased to ≥ 95% in all SV length categories, SV calls of this combination were kept for the particular SV type and were combined with a logical or with those of the first step. The threshold of ≥ 95% precision was used to reduce the number of FP SV calls to a reasonable level.

Detection of SV, SNV, and INDELs in the barley genome

Genetic material and sequencing

Our study was based on 23 spring barley inbreds (Weisweiler et al. 2019) that were selected out of a worldwide collection of 224 inbreds (Haseneyer et al. 2010) (Supplementary Table S6) using the MSTRAT algorithm (Gouesnard 2001). These inbreds are the parents of the double round robin population (Casale et al. 2022). Paired-end sequencing libraries with an insert size of 425 bp were sequenced (2x150 bp) to a ~25x coverage on the Illumina HiSeqX platform by Novogene Corporation Inc. (Sacramento, USA).

SV, INDELs, and SNV detection

The quality of the raw reads was checked by fastqc. Reads were adapter- and quality-trimmed using Trimmomatic (version 0.39) (Bolger et al. 2014). The trimmed reads were mapped to the Morex reference sequence v3 (Mascher et al. 2021) using BWA-MEM. PCR-duplicates were removed using PICARD (version 2.22.0).

Based on the results of the benchmarking of different SV callers using simulated data, the results of specific SV callers were combined as explained above. The final set of deletions for each inbred were those that were identified by Manta | GRIDSS | Pindel | Delly | (Lumpy & NGSEP) where homozygous-reference (0/0) and heterozygous variant (0/1) calls were removed. Additionally, deletions annotated by Pindel as RPL were removed. In analogy, the duplications were identified by Manta | GRIDSS | Pindel | (Delly & Lumpy). Insertions of the SV length category A were identified by Manta | GRIDSS | Delly, where insertions of the SV length categories B-E were called using Manta. Inversions were identified by Manta | GRIDSS | Pindel. Translocations were called from pairs of break points identified by Manta | GRIDSS | (Delly & Lumpy). INDELs were detected by GATK's HaplotypeCaller | GRIDSS | Pindel where homozygous-reference (0/0) and heterozygous variant (0/1) calls were discarded. SV which were located in a region of the reference sequence, where the sequence only consists of N's, were excluded. For genome regions, where break points of different SV overlapped or were inconsistent in the same inbred, only the smallest SV was considered.

The number of false positives could be increased by detecting large SV clusters; therefore, SV clusters larger than 1 Mb were not considered in our study. The SV of the 23 inbreds were grouped together to SV clusters based on the similarity of sizes and the position in the genome according to the following procedure. The distance from a SV to the next SV in such a SV cluster had to be smaller than 20 bp for the SV length category A and 50 bp for the SV length category B - E and the difference of the two break points had to be smaller than 10 or 50 bp as described above. SV with a larger difference between break points were kept as separate SV and SV clustering was pursuing. Each SV cluster was genotyped across the examined 23 barley inbreds.

SNV and INDELs were called using GATK. First, GATK's HaplotypeCaller was used in single sample GVCF mode, afterward GATK's CombineGVCFs was used to combine the SNV across the 23 inbreds. Combined SNV were genotyped using GATK's GenotypeGVCFs. SNV were filtered using GATK's VariantFiltration where variants below the following filtering thresholds were removed: QD < 2.0; QUAL < 30.0; SOR > 3.0; FS > 60.0; MQ < 40.0; MQRankSum < −12.5; ReadPosRankSum < −8.0. Heterozygosity of SNV for each genotype was low (1.0–1.7%) and therefore such SNV were not discarded to avoid removing true positives.

PCR validation of SV

A total of 25 of the detected SV were targeted for validation by PCR amplification of genome regions of and around the SV in Morex and Unumli-Arpa. This included six SV length category A deletions, five SV length category A insertions, six SV length category B deletions and eight SV length category C-E deletions. In order to determine the SV allele, we required the amplification of two differently sized fragments in the two inbreds. For each SV, a regular primer pair was created with the position defined by the validation strategy (Supplementary Fig. S1). If needed, a second right primer was added to the PCR reaction. The primers were designed using Primer3 (Untergasser et al. 2012) and Blast+ (Camacho et al. 2009).

Plant material was sampled for the PCR validation from adult plants and seedlings grown under controlled conditions. DNA was extracted from 100 mg frozen plant material using the DNeasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The PCR reaction mixture contained in a final volume of 20 µL: 0.2 mM dNTP, Fw/Rev Primer 0.5 µM, 50 ng DNA, 1.5 U/µL DreamTaq DNA Polymerase (Thermo Fischer Scientific, USA), Polymerase-Buffer 1X and water. Amplified fragments were separated by gel electrophoresis and the validation success was determined by comparing the PCR product sizes with the calculated values based on the SV detection.

Location of SV clusters

SV clusters were classified and annotated based on their location in the genome, their distance relative to genes, or other genomic features. SV clusters were grouped into four gene-associated and one intergenic SV cluster categories: 5 kb upstream/downstream gene-associated SV clusters were located in the 5 kb region from the 3′- or 5′- end of a gene. Intron and exon gene-associated SV clusters were located in the gene sequence, where the genic sequence was separated into intronic and exonic sequences. SV clusters which were not located in the four gene-associated SV cluster categories were determined as intergenic SV clusters. A gene-associated SV cluster could be classified in more than one category if its sequence covers several genomic features.

To check if the detected SV clusters were transposable elements, the genomic positions of SV clusters were compared to the transposable elements annotation file of the Morex reference sequence v3 (Mascher et al. 2021). Deletions, duplications, inversions, INDELs, and insertions with known length were annotated as transposable elements if the reciprocal overlap was ≥ 80% (Fuentes et al. 2019). Insertions with unknown length were classified as transposable elements if the detected break point of the insertion was inside the transposable element sequence. Translocations were classified as transposable element, if at least one of the two break points was located inside a transposable element sequence.

SV hotspots were identified using the following procedure: The average number of SV clusters in non-overlapping 1 Mb windows across each of the seven chromosomes was determined. Using this number, we calculated for each window based on the Poisson distribution the expected number of SV clusters. Windows with more SV clusters than the Q_{99} of the expected Poisson distribution were designated as SV hotspots (Guan et al. 2021).

Population genetic analyses

Linkage disequilibrium (LD) measured as r^2 (Hill and Robertson 1968) was calculated between each SV type and linked SNV. Nucleotide diversity (π) was calculated in 100 kb windows along the seven chromosomes separately for SV clusters (deletions, insertions, duplications, inversions) and SNV using vcftools (version 0.1.17) (Danecek et al. 2011).

SV clusters and gene expression

SV clusters which were assigned into one of the gene-associated SV categories, namely 5 kb up- or downstream, introns, and exons, were associated with the genome-wide gene expression of the 23 barley inbreds. Gene expression for the

seedling tissue measured as fragments per kilobase of exon model per million fragments mapped was available for all inbreds from an earlier study (Weisweiler et al. 2019). This information was the basis of a principal component analysis. For all gene-associated SV clusters with a minor allele frequency (MAF) > 0.15, Pearson's correlation coefficient with the first three principal components was estimated, where the presence and absence of SV clusters were used as metric character. This analysis was performed to examine the association between SV clusters and genome-wide gene expression (Liu et al. 2020b). A permutation procedure with 1,000 iterations was used to test the mean absolute values of the correlations for their significance. In addition to this evaluation of the effect of SV clusters on the genome-wide gene expression level, we also examined the significance of the effect of gene-associated SV clusters with a MAF > 0.15 on the expression of individual genes. In order to do so, the mixed linear model with population structure and kinship matrix (PK model) (Stich et al. 2008) was used. The population structure matrix consisted of the first two principal components calculated from 133,566 SNV and INDELs derived from mRNA sequencing (Weisweiler et al. 2019). From the same information, the kinship matrix was calculated as described by Endelman and Jannink (2012).

Assessment of phenotypic traits

For the assessment of phenotypic traits under field conditions, the 23 inbreds were planted as replicated checks in an experiment laid out as an augmented row-column design. The experiment was performed in seven environments (Cologne from 2017 to 2019, Mechernich and Quedlinburg from 2018 to 2019) in Germany in which the checks were replicated multiple times per environment. For each environment, seven phenotypic traits were assessed. Heading time (HT) was recorded as days after planting, leaf angle (LA) was scored on a scale from 1 (erect) to 9 (very flat) on four-week-old plants, and plant height (PH, cm) was measured after heading in Cologne and Mechernich. Seed area (SA, mm^2), seed length (SL, mm), seed width (SW, mm), and thousand grain weight (TGW, g) were measured based on full-filled grains from Cologne (2017–2019) and Quedlinburg (2018) by using MARVIN seed analyzer (GTA Sensorik, Neubrandenburg, Germany).

Prediction of phenotypes

Each of the phenotypic traits was analyzed across the environments using the following mixed model:

$$y_{ijk} = \mu + E_j + G_i + (G \times E)_{ij} + \varepsilon_{ijk}, \quad (4)$$

where y_{ijk} was the observed phenotypic value for the i^{th} genotype at the j^{th} environment within the k^{th} replication; μ the general mean, G_i the effect of the i^{th} inbred, E_j the effect of the j^{th} environment, $(G \times E)_{ij}$ the interaction between the i^{th} inbred and the j^{th} environment, and ε_{ijk} the random error. This allowed to estimate adjusted entry means for all inbreds.

The performance to predict the adjusted entry means of each barley inbred for each trait using different types of predictors: (1) single nucleotide polymorphism (SNP) array, which was generated by genotyping the 23 inbreds using the Illumina 50K barley SNP array (Bayer et al. 2017), (2) gene expression (3) SNV & INDELs, (3a) SNV, (3b) INDELs, (4) SV clusters, (4a) deletions, (4b) duplications, (4c) insertions, (4d) inversions, (4e) translocations, was compared based on genomic best linear unbiased prediction (GBLUP) (VanRaden 2008).

For each predictor, the monomorphic features and the features with missing rates > 0.2 and identical information were discarded. **W** was defined as a matrix of feature measurement for the respective predictor. The dimensions of **W** were the number of barley inbreds ($n = 23$) times the number of features in the corresponding predictor (m) ($m_{SNP\ \text{array}} = 38,025$, $m_{\text{gene expression}} = 67,844$, $m_{\text{SNV\&INDELs}} = 3,025,217$, $m_{\text{SNV}} = 2,338,565$, $m_{\text{INDELs}} = 686,918$, $m_{\text{SVclusters}} = 458,330$, $m_{\text{deletions}} = 183,219$, $m_{\text{duplications}} = 93,073$, $m_{\text{insertions}} = 70,143$, $m_{\text{inversions}} = 6,582$, $m_{\text{translocations}} = 105,313$). The additive relationship matrix **G** was defined as $\mathbf{G} = \frac{\boldsymbol{W}^*\boldsymbol{W}^{*T}}{m}$, where $\boldsymbol{W}^*$ was a matrix of feature measurement for the respective predictor, whose columns are centered and standardized to unit variance of $\boldsymbol{W}$, and $\boldsymbol{W}^{*T}$ was the transpose of $\boldsymbol{W}^*$.

Furthermore, to investigate the performance of a joined weighted relationship matrix (Schrag et al. 2018) to predict phenotypic variation, the three **G** matrices in GBLUP model of the three predictors, SNV &INDELs, gene expression, and SV clusters, were weighted and summed up to one joined weighted relationship matrix. A grid search, varying any weight (w) from 0 to 1 in increments of 0.1, resulted in 66 different combinations of joined weighted relationship matrix, where the summation of three weights in each combination must be equal to 1.

Fivefold cross-validation was used to assess the model performance. Prediction abilities were obtained by calculating Pearson's correlations between observed (y) and predicted ($\hat{y}$) adjusted entry means in the validation set of each fold. The median prediction ability across the five folds within each replicate was calculated and the median of the median across the 200 replicates was used for further analyses.

Results

Precision and sensitivity of SV callers

Six tools (Table 1) which call SV based on short-read sequencing data were evaluated with respect to their precision and sensitivity to detect five different SV types (deletions, insertions, duplications, inversions, and translocations) in five SV length categories (A: 50–300 bp; B: 0.3–5 kb; C: 5–50 kb; D: 50–250 kb; E: 0.25–1 Mb) using computer simulations. The precision of Delly, Manta, GRIDSS, and Pindel to detect deletions of all five SV length categories based on 25x sequencing coverage ranged from 97.8–100.0%, whereas the precision of Lumpy and NGSEP was lower with values between 75.0 and 89.8% (Table 2). The sensitivity of NGSEP was with 78.6–87.5% the highest but that of Manta was with 79.7–81.1% only slightly lower. We evaluated various combinations of SV callers and observed for the combination of Manta | GRIDSS | Pindel | Delly | (Lumpy & NGSEP) an increase of the sensitivity to detect deletions compared to the single SV callers up to a final of 89.0% without decreasing the precision considerably (99.1%).

Manta was the only SV caller which allowed the detection of insertions of all SV length categories with precision values as high as 99.8–100.0%. The combination of Manta | GRIDSS | Delly for the SV length category A has shown a high sensitivity (88.4%) and precision (99.8%). This combination was therefore used for the detection of insertions of SV length category A in further analyses.

The sensitivity of the SV callers Delly, Manta, Lumpy, and GRIDSS to detect duplications of the SV length category A was with values from 28.2 to 39.4% very low. In contrast, Pindel could detect these duplications with a sensitivity of 75.7%. For the other SV length categories, the combination of Manta | GRIDSS | Pindel could increase the sensitivity to detect duplications by 2–7% compared to using a single SV caller while the precision ranged between 97.6 and 99.3%.

The performance of Lumpy and NGSEP to detect inversions reached precision values of 81.5–98.5% and sensitivity values of 66.1–80.0% that were on the same low level as for deletions. Delly performed well for detecting inversions in SV length categories B to D, but for E and especially for A, the performance was lower compared to that of the other SV callers. Overall, Pindel was the only SV caller with a combination of both, high precision and sensitivity to detect inversions. These precision and sensitivity values could be further improved across all SV length categories by combining the calls of Pindel with that of Manta | GRIDSS (Table 2).

The combination of GRIDSS | Pindel | GATK increased the sensitivity to detect INDELs (2–49 bp) by 3% compared to using the single callers (Supplementary Table S1). With 6%, an even higher difference for the sensitivity to detect translocations was observed between the combination of Manta | GRIDSS | (Delly & Lumpy) and single callers.

In a next step, different sequencing coverages from 1.5x to 65x were simulated and the performance of the best combination of SV callers for each of the SV types was compared to their performance with 25x sequencing coverage (Supplementary Fig. S1). For deletions, the F1-score, which is harmonic mean of the precision and sensitivity, for 65x sequencing coverage was ~2% higher than for 25x sequencing coverage. Only marginal differences were observed between the F1-score of 65x or 25x sequencing coverage for calling duplications and inversions. Interestingly, the F1-score for calling translocations and insertions was with 2% and 9%, respectively, higher in the scenario with 25x than with 65x sequencing coverage. For 12.5x sequencing coverage, the F1-score was still on an high level with values > 80% for each SV type (Supplementary Fig. S2). With a further reduced sequencing coverage, the F1-score also decreased. Finally, the performance of our pipeline to detect SV was evaluated based on 14x and 25x linked-read sequencing data. For all SV types and SV length categories, with the exception of deletions and duplications in SV length category D and A, respectively, the F1-score was 2–7% higher based on Illumina sequencing data than based on linked-read sequencing data.

Table 1 Properties of structural variant (SV) callers for short-read sequencing that were compared in our study, where split reads (SR), paired-end reads (PE), read depth (RD), and local alignments (LA) are the underlying detection principles

SV caller	Detection principle				Deletion	Insertion		Inversion	Duplication	Translocation
	SR	PE	RD	LA		≤500bp	>500bp			
Pindel[1]	x				x	x	x	x	x	
Delly[2]	x	x			x	x		x	x	x
Lumpy[3]	x	x	x		x			x	x	x
Manta[4]	x	x		x	x	x	x	x	x	x
GRIDSS[5]	x	x		x	x	x		x	x	x
NGSEP[6]			x		x	x	x	x		

[1]Ye et al. (2009), [2]Rausch et al. (2012), [3]Layer et al. (2014), [4]Chen et al. (2016), [5]Cameron et al. (2017), [6] Duitama et al. (2014)

Table 2 Sensitivity/precision of structural variant (SV) callers and combinations of them (for details see Material & Methods) to detect deletions, insertions, duplications, and inversions of the SV length categories A (50–300 bp), B (0.3–5 kb), C (5–50 kb), D (50–250 kb), and E (0.25 – 1 Mb)

SV caller	SV length category				
	A	B	C	D	E
	Deletions				
Delly	58.1/97.8	76.2/99.4	72.5/99.3	72.4/100.0	75.0/100.0
Manta	79.7/100.0	81.1/99.8	79.9/99.6	79.7/99.4	81.0/100.0
Lumpy	60.0/78.1	70.5/86.5	66.8/85.6	62.5/79.0	64.3/80.6
GRIDSS	79.0/99.5	80.7/99.9	77.8/99.9	78.1/100.0	77.4/100.0
Pindel	87.4/99.9	68.4/99.7	83.6/99.4	80.2/100.0	67.9/100.0
NGSEP	84.1/87.3	83.1/83.4	83.5/82.2	87.5/89.8	78.6/75.0
Combination	89.0/99.1	86.9/99.4	86.7/99.2	86.5/99.4	86.9/100.0
	Insertions				
Delly	3.4/100.0				
Manta	88.4/99.8	74.1/100.0	72.1/100.0	72.5/100.0	75.0/100.0
GRIDSS	45.5/100.0				
Pindel	6.6/93.0				
NGSEP	64.1/59.2	26.8/29.6	35.5/40.5	30.5/32.1	26.0/26.5
Combination	88.4/99.8	74.1/100.0	72.1/100.0	72.5/100.0	75.0/100.0
	Duplications				
Delly	28.2/99.0	75.1/96.8	74.7/95.4	75.3/97.2	71.7/91.7
Manta	39.0/99.5	80.5/99.8	82.7/99.8	83.9/98.7	82.6/97.4
Lumpy	31.5/98.4	67.9/84.8	67.7/82.6	68.3/81.9	65.2/80.0
GRIDSS	39.4/99.8	80.0/100.0	80.0/100.0	83.3/100.0	79.4/100.0
Pindel	75.7/98.1	57.8/99.0	88.1/99.8	83.9/99.4	73.9/100.0
Combination	75.8/98.1	87.3/99.1	90.8/99.3	89.8/98.2	89.1/97.6
	Inversions				
Delly	49.7/70.4	84.6/99.2	85.5/99.4	82.6/99.4	78.2/98.6
Manta	77.0/99.0	87.0/99.9	87.3/99.9	90.0/100.0	82.8/100.0
Lumpy	66.1/88.5	76.8/96.2	75.3/97.4	77.4/94.8	74.7/98.5
GRIDSS	76.9/99.1	86.9/99.8	85.2/99.9	87.9/100.0	82.8/100.0
Pindel	83.5/99.2	90.7/99.9	90.2/99.9	89.0/100.0	77.0/100.0
NGSEP	0.0/0.0	75.7/87.9	75.3/81.5	80.0/85.4	77.0/88.2
Combination	88.4/98.1	91.5/99.8	90.9/99.8	93.2/100.0	85.1/100.0

SV clusters across the 23 parental inbreds of the double round robin population

Across the 23 barley inbreds that are the parents of a new resource for mapping natural phenotypic variation, the double round robin population, we detected 458,671 SV clusters using the best combination of SV callers (Table 3). These comprised 183,489 deletions, 70,197 insertions, 93,079 duplications, 6,583 inversions, and 105,323 translocations. Additionally, 6,381,352 INDELs were detected

Table 3 Summary of detected structural variants (SV) and small insertions and deletions (2–49 bp, INDELs) across 23 diverse barley inbreds, where MAF was the minor allele frequency, and TE were SV clusters which were annotated as transposable elements in the Morex reference sequence v3

SV type	Number of SV calls	Number of SV clusters		
			MAF > 0.05	TE
Deletions	714,867	183,489	78,823	16,846
Insertions	241,522	70,197	29,672	279 (17,718)[1]
Duplications	195,710	93,079	58,793	6,608
Inversions	14,961	6,583	4,116	92
Translocations	251,956	105,323	61,572	0 (54,258)[1]
INDELs	29,637,520	6,381,352	4,134,064	21

[1]Because of missing endpoint information no reciprocal overlap criterion applied

across the seven chromosomes. The proportion of SV clusters which were annotated as transposable elements varied from 1.4% for inversions to 51.5% for translocations.

We performed a PCR-based validation for detected deletions and insertions (Supplementary Table S2, Supplementary Fig. S3). Six out of six deletions and five out of five insertions up to 0.3 kb could be validated (Supplementary Fig. S4). Additionally, we could validate eight out of eleven deletions between 0.3 and 460 kb (Supplementary Fig. S5), where for the three not validated deletions, the expected fragments were not observed in the non-reference parental inbred.

The number of SV clusters present per inbred ranged from less than 40,000 to more than 80,000 (Fig. 1A). We observed no significant ($P > 0.05$) correlation between the sequencing coverage, calculated based on raw, trimmed, and mapped reads, of each inbred as well as the number of detected SV clusters in the corresponding inbred. A two-sided t-test resulted in no significant ($P > 0.05$) association between the number of SV clusters of an inbred and the spike morphology as well as the landrace versus variety status of the inbreds. In contrast, principal component analyses based on the presence/absence matrices of the SV clusters revealed a clustering of inbreds by spike morphology, geographical origin, and landrace vs. variety status (Supplementary Fig. S6).

Out of the 458,671 SV clusters, 50.6% (232,071) appeared in only one of the 23 inbreds, whereas 19.7% (90,256) were detected in at least five inbreds (Fig. 1B, Supplementary Fig. S7). Additional analyses revealed a significant although weak negative correlation ($r = -0.06681$, $P = 2.07\text{x}10^{-314}$) between the length of a SV cluster and its MAF. The average MAF of SV clusters with a length of 250 kb to 1 Mb and of 50–250 kb was 0.08, respectively, while that of SV clusters with a length of 50 bp–50 kb was 0.13 (Supplementary Fig. S8). SV clusters annotated as transposable elements had a shorter average length of 5,853 bp and a higher MAF of 0.16 compared to SV clusters that were not annotated as transposable elements (10,605 bp, 0.12). Deletions and insertions of the SV length category A were the most common detected SV clusters with a fraction of 41.7 and 48.4%, respectively (Supplementary Table S3). In contrast, for duplications, the

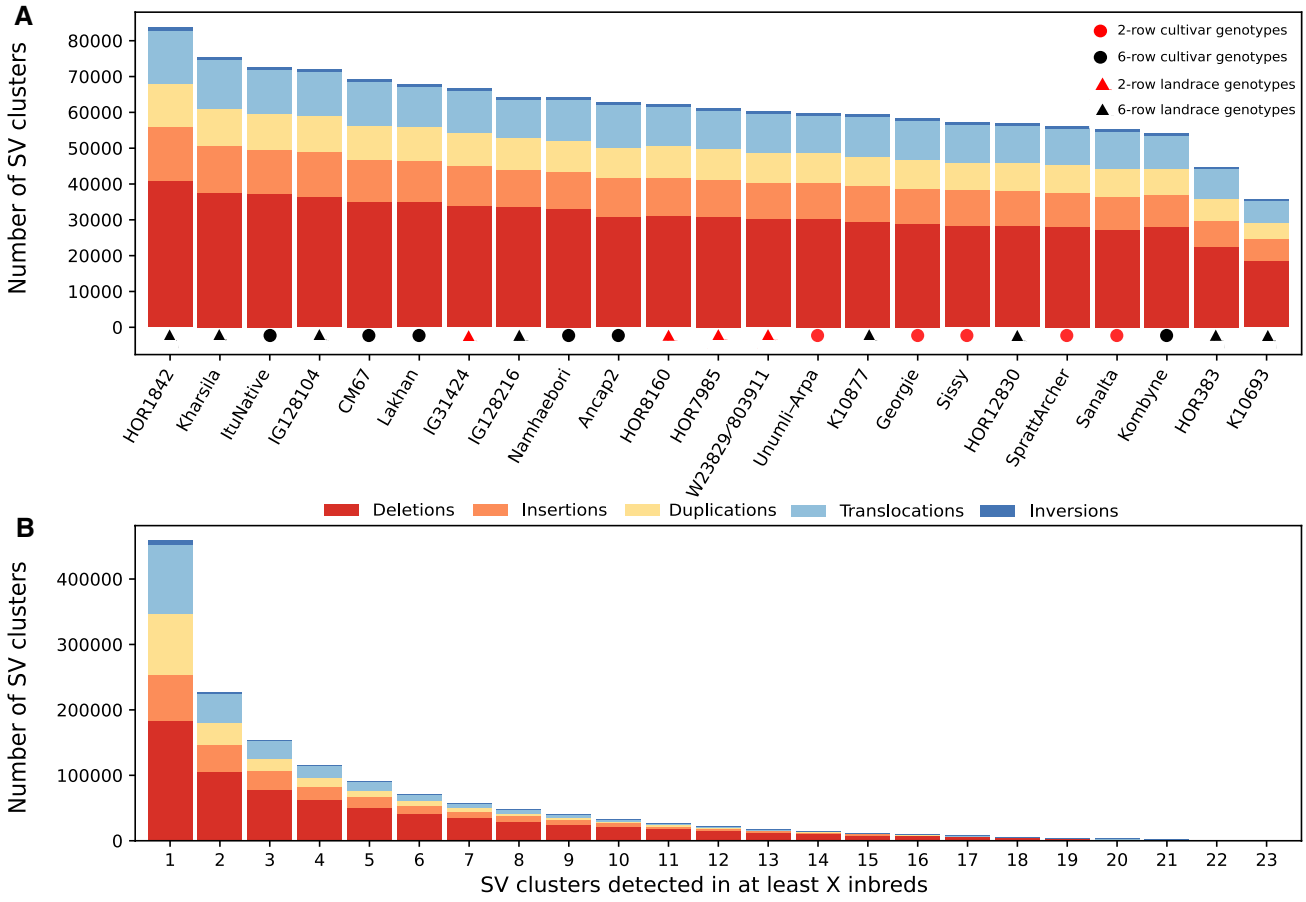


Fig. 1 Stacked bar graph of the number of different types of structural variant (SV) clusters detected in the 23 inbreds (**A**) and SV clusters which were detected in at least the given number of the inbreds (**B**)

Theoretical and Applied Genetics (2022) 135:3511–3529

largest fraction were that for SV clusters of the SV length category C (55.9%). The average MAF of the individual SV types was the highest for insertions with 0.17, followed by deletions, inversions, translocations, and duplications with values of 0.14, 0.11, 0.10, and 0.10, respectively.

Characterization of the SV clusters

After examining the length of the detected SV clusters and their presence in the 23 barley inbreds, we investigated the distribution of the SV clusters across the barley genome. We observed a significant correlation ($r = 0.5653$, $P < 0.01$) of nucleotide diversity (π) of SV clusters and SNV, measured in 100 kb windows along the seven chromosomes (Supplementary Fig. S9). The SV clusters were predominantly present distal of pericentromeric regions. In contrast to SNV, the frequency of all SV types, and especially that of duplications, increased in centromeric regions (Fig. 2). For all centromeres, a significantly ($P < 0.01$) higher number of SV clusters was observed compared to what is expected based on a Poisson distribution and, thus, were designated as SV hotspots. The proportion of SV clusters in pericentromeric regions was with 14.5% considerably lower compared to what is expected based on the physical length of these regions (25.7%). Only 4.5% of all detected SV hotspots were observed in pericentromeric regions.

We also examined if SV clusters provide additional genetic information compared to that of closely linked SNV. To do so, we determined the extent of LD between each SV cluster and SNV located within 1 kb and compared this with the extent of LD between the closest SNV to the SV cluster and the SNV within 1 kb. Across the different SV types, 33.7–74.3% have at least one SNV within 1 kb that showed an $r^2 \geq 0.6$ (Supplementary Table S4). In contrast, 89.2–89.9% of SNV that are closest to the SV cluster showed an $r^2 \geq 0.6$ to another SNV within 1 kb.

In the next step, we examined the presence of SV clusters relative to the position of genes. The highest proportion of SV clusters (~60%) was located in intergenic regions of the genome (Fig. 3). The second largest fraction (~30%) of SV clusters was present in the 5 kb up- or downstream regions of genes, which is considerably higher compared to that of INDELs (~17%) and SNV (~16%). Within the group of SV clusters that were 5 kb up- or downstream to genes, a particularly high fraction were inversions. On average across all SV types, about 10% of SV clusters were located in introns and exons, with inversions being the exception again, showing a considerably higher rate.

The enrichment of SV clusters proximal to genes lead us to assess their physical distance relative to the transcription start site (TSS) of the closest genes and compare this to SNV. The number of SV clusters at the TSS was approximately 10% lower than 5kb upstream of the TSS (Fig. 4). A similar trend was observed for the 5kb downstream regions (~7%). In comparison, the absolute number of SNV around the TSS was more than ten times lower than the number of SV clusters. With the exception of a distinct peak at position two downstream of the TSS, the number of SNV around the TSS followed the same trends as described for the SV clusters above.

Association of SV clusters with gene expression

We tested if the SV clusters could be associated with the genome-wide gene expression differences of the 23 inbreds. As a first step, a principal component analysis of the gene expression matrix, which included all genes and inbreds, was performed. The loadings of all 23 inbreds on principal component (PC) 1 explained 19.7% of the gene expression variation and were correlated with the presence/absence status of all inbreds for each gene-associated SV cluster. The average absolute correlation coefficient of gene-associated SV clusters and the PC1 of gene expression was 0.17 and higher than the Q_{95} of the coefficient observed for randomized presence/absence pattern and the PC1 (Supplementary Fig. S10, Supplementary Fig. S11). Similar observations were made for the association of gene-associated SV clusters with PC2 and PC3 of 0.17 and 0.19, respectively, for the above-mentioned gene expression matrix (Supplementary Fig. S12). In addition, we investigated a possible association between SV clusters and gene expression on the basis of individual genes. For a total of 1,976 out of 21,140 gene-associated SV clusters a significant ($P < 0.05$) association with the gene expression of the associated gene was observed (Fig. 5).

Prediction of phenotypic variation from SV clusters

The prediction ability of seven quantitative phenotypic traits using SV clusters as well as SNV from a SNP array, genome-wide gene expression information, SNV and INDELs (SNV & INDELs) were examined as predictors through five-fold cross-validation. The median prediction ability across all traits ranged from 0.509 to 0.648. The SV clusters had the highest prediction power, followed by SNV & INDELs, SNP array, and gene expression in decreasing order (Fig. 6). Compared to these differences, those among the median prediction abilities of the different SV types were small. The highest prediction ability was observed for insertions and the lowest for inversions. We also evaluated the possibility to combine SNV and INDELs with gene expression and SV cluster information using different weights to increase the prediction ability (Supplementary Fig. S13). The mean of the optimal weight across the seven traits was highest for gene expression (0.41) and lowest for SV clusters (0.23) (Supplementary Table S5).

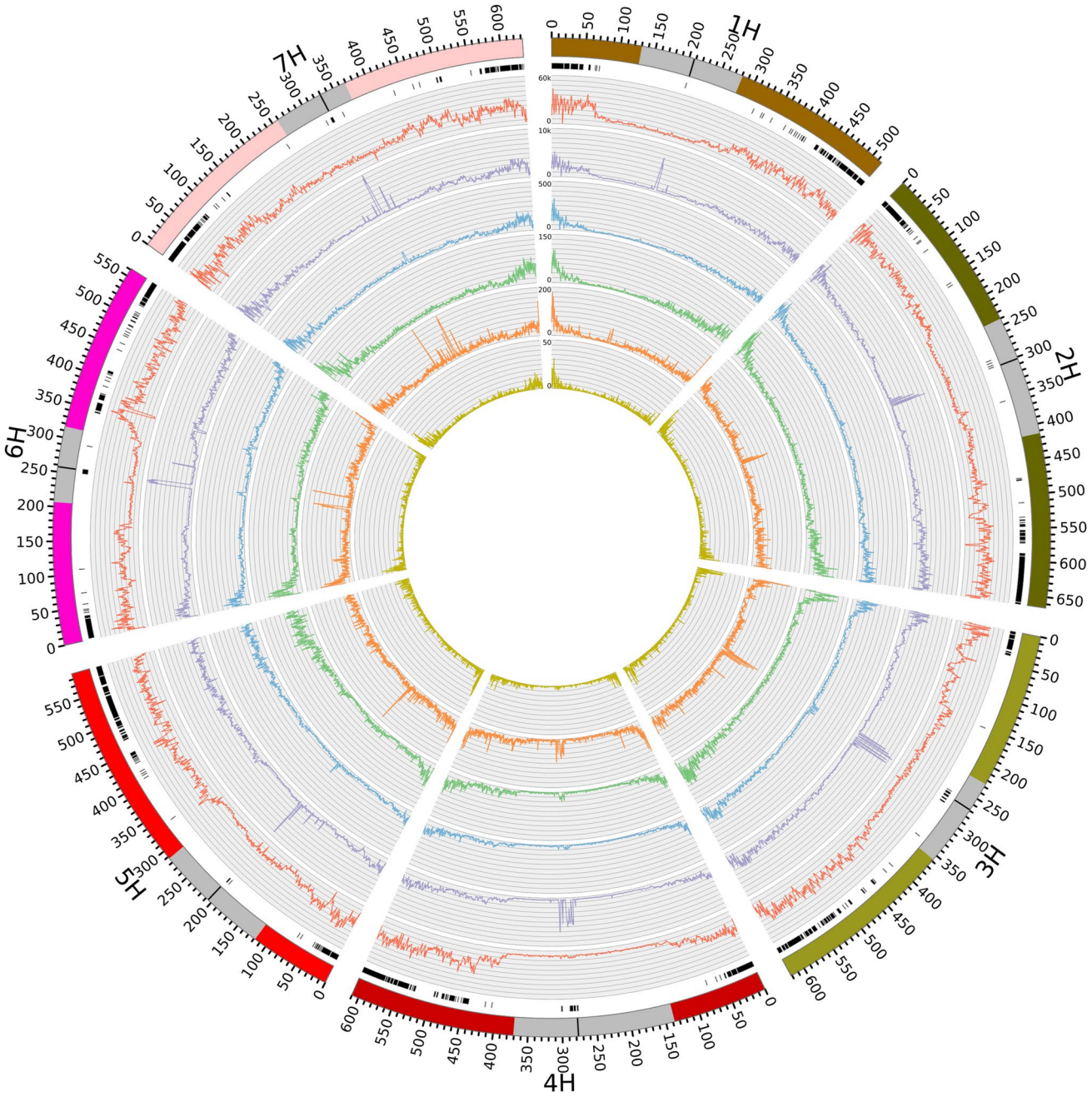


Fig. 2 Distribution of genomic variants among 23 barley inbreds across the seven chromosomes. The outermost circle denotes the chromosome number, the physical position, and as gray bar the pericentromeric regions (Casale et al. 2022) plus the centromeres (black) according to the Morex reference sequence v3. The next inner circles report the SV cluster hotspots (black bars), frequencies of single-nucleotide variants (red), small insertions and deletions (2–49 bp, INDELs, purple), deletions (blue), insertions (green), duplications (orange), and inversions (yellow) which were detected among the 23 inbreds (color figure online)

Discussion

The improvements to sequencing technologies made SV detection in large genomes possible (Della Coletta et al. 2021). Despite these advances, the relative high cost of third compared to second generation sequencing makes the former less affordable and scalable for many research groups. This fact is particularly strong if genotypes have to be analyzed. We therefore used computer simulations to study the precision and sensitivity of SV detection based on different sequencing coverages of short-read sequencing data in the model cereal barley. We also evaluated

Theoretical and Applied Genetics (2022) 135:3511–3529

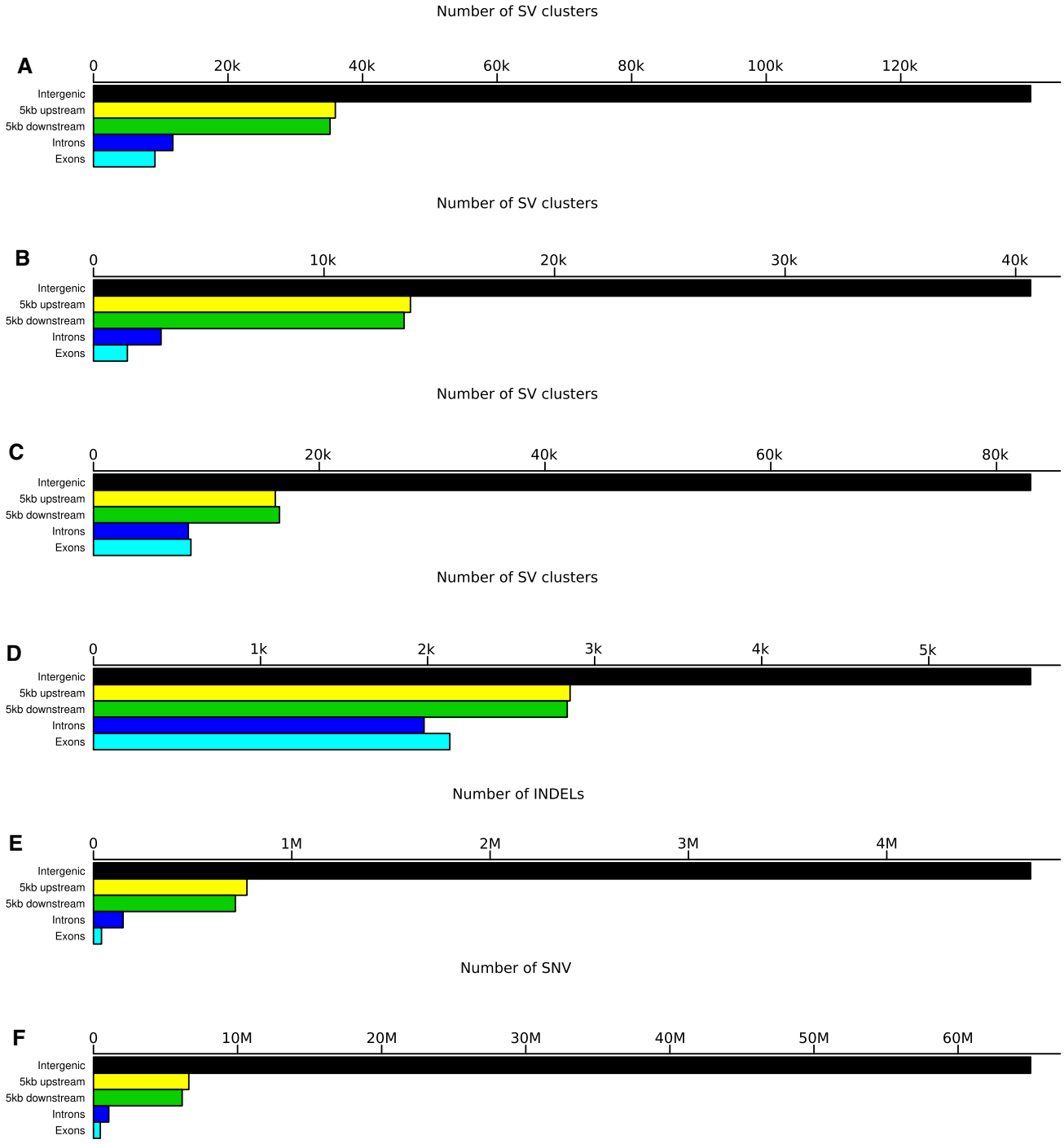


Fig. 3 The occurrence of deletions (**A**), insertions (**B**), duplications (**C**), inversions (**D**), small insertions and deletions (2 –0 49 bp, INDELs, **E**), and single-nucleotide variants (SNV) (**F**) in five genomic regions

whether linked-read sequencing offered by BGI (Wang et al. 2019) or formerly 10x Genomics (Weisenfeld et al. 2017) is advantageous for SV detection compared to classical Illumina sequencing.

Limitations of our study

In our study, the different SV types were always determined in comparison against one reference sequence. The number of insertions present in this reference inbred determines the number of detected deletions and vice versa. However, this

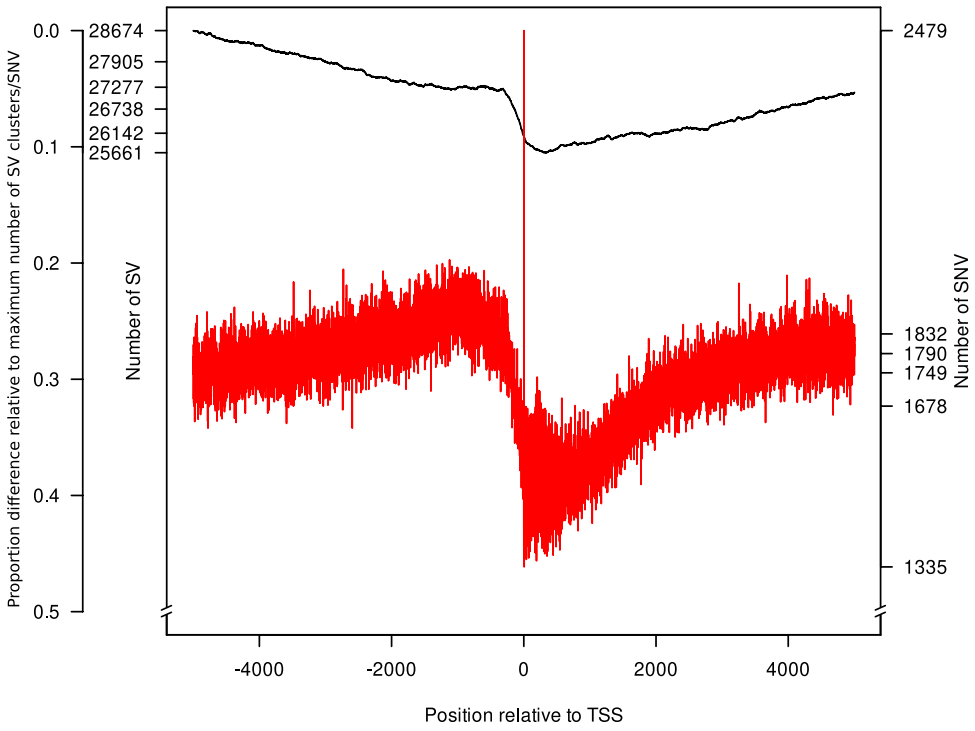


Fig. 4 Distribution of structural variant (SV) clusters (black) and single-nucleotide variants (SNV, red) among 23 barley inbreds relative to the transcription start site (TSS) of a gene (x-axis). SV clusters and SNV were counted for every position from 5kb up- and downstream around the TSS of all genes (y-axes). As third y-axis, the proportion difference relative to the maximum number of SV clusters/SNV is illustrated (color figure online)

is just a matter of nomenclature. Additionally, the usage of short-read sequencing data and only one reference sequence could lead to detect false positive SV calls, due to differences in the mapping efficiency of the evaluated inbreds due to differences in relatedness. In our study, however, the average mapping quality for the 23 inbreds was high and varied only moderately between 41 and 46. Therefore, the influence of the relatedness should be weak. However, this aspect should be considered when interpreting the SV data set.

Precision and sensitivity to detect SV in complex cereal genomes using short-read sequencing data are high

The costs for creating linked-read sequencing libraries is considerably higher compared to that of classical Illumina libraries. Taking this cost difference into account, a fair comparison of precision and sensitivity to detect SV is between 25x Illumina and 14x linked-reads. However, even when directly compared at equal (25x) sequencing coverage, the F1-score, which is the harmonic mean of the precision and sensitivity, on average across all SV types and SV length categories was higher for Illumina compared to linked-reads (Supplementary Fig. S1). One reason might be that the SV callers used in our study do not fully exploit linked-read data. In our study, linked-read information was only used to improve the mapping against the reference genome (Marks et al. 2019). More recently, SV callers have been described that exploit linked information of linked-read data as VALOR2 (Karaoğlanoğlu et al. 2020) or LEVIATHAN (Morisse et al. 2021). However, the SV callers that were available at the time the simulations were performed had a very limited spectrum of SV types and SV length categories they could detect, e.g., LongRanger wgs (Zheng et al. 2016) and NAIBR (Elyanow et al. 2018). In addition, we have observed for these SV callers in first pilot simulations considerably lower values for precision and sensitivity to detect SV compared to the classical short-read SV callers. Therefore, only short-read SV callers were evaluated in detail.

One further aspect that we examined was the influence of the sequencing coverage on sensitivity and precision of SV detection. Only a marginal difference between the F1-scores of the best combination of SV callers for a 25x vs. 65x Illumina sequencing coverage was observed (Supplementary Fig. S1). In addition, for some SV length categories, the F1-score for 25x compared to 65x sequencing coverage was actually higher. A possible explanation for this observation may be that a higher sequencing coverage can lead to an increased number of spuriously aligned reads (Kosugi et al. 2019). These reads can lead to an increased rate of false positive SV detection (Gong et al. 2021). Our result suggests that for homozygous genomes, Illumina short-read sequencing coverage of 25x is sufficient to detect SV with a high precision and sensitivity. We therefore made use of this sequencing coverage not only for further simulations but also to re-sequence the 23 barley inbreds of our study.

In addition, we also tested if a lower sequencing coverage could be used for SV detection to reduce the cost for

Theoretical and Applied Genetics (2022) 135:3511–3529

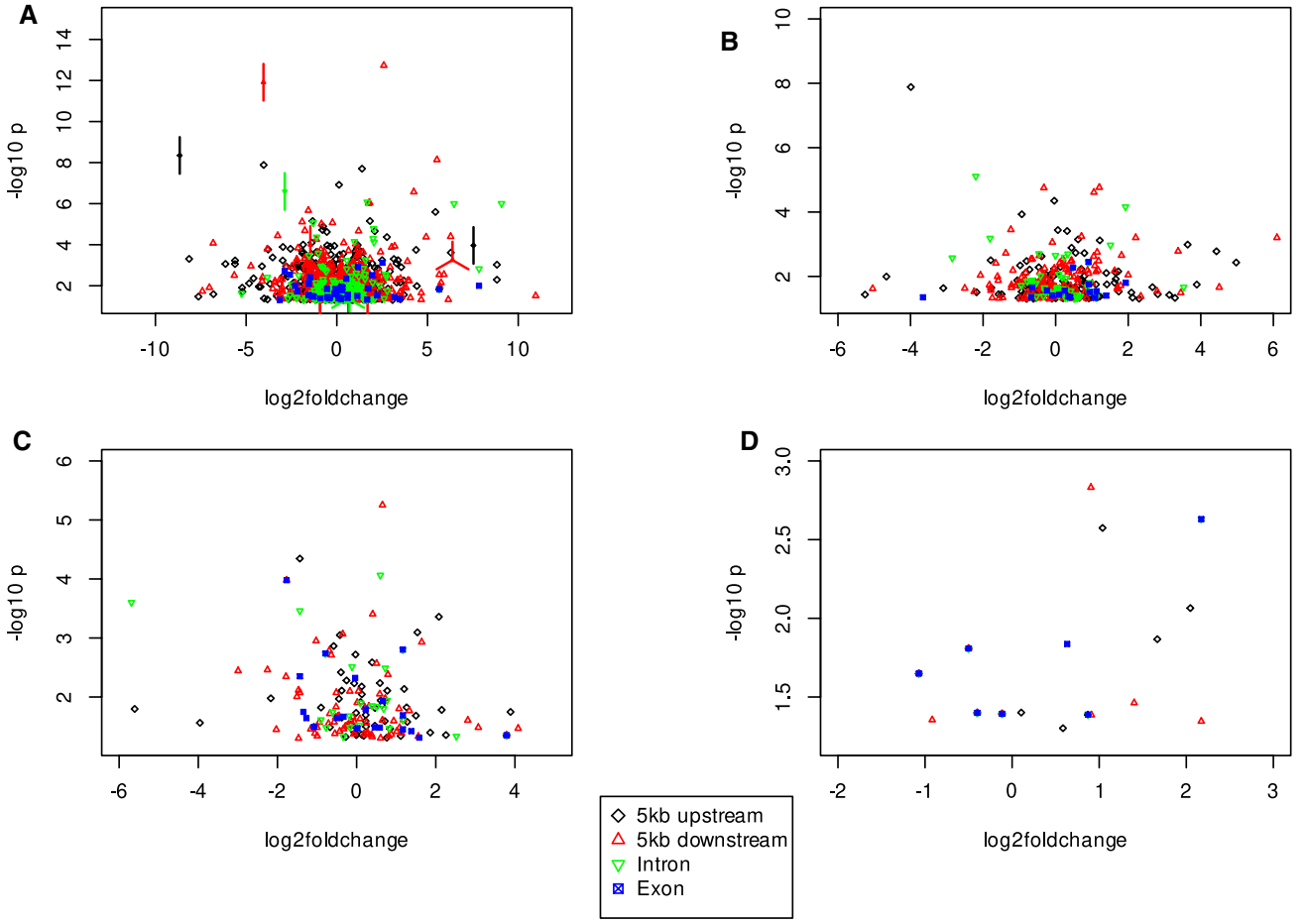


Fig. 5 Association of gene-associated (for details see Material & Methods) deletions (**A**), insertions (**B**), duplications (**C**), and inversions (**D**) with a minor allele frequency > 0.15 with the expression of individual genes assessed using the PK mixed linear model. The gene-associated structural variant (SV) clusters were classified based on their occurrence relative to genes in 5kb up- or downstream, introns, and exons. Values of SV clusters with the same coordinates are illustrated as points with edges, where each edge represents one SV cluster

sequencing further. We observed lower F-scores for all SV types using a sequencing coverage of 12.5x than for 25x (Supplementary Fig. S2). However, the F1-score was still > 80% for all SV types suggesting that even a sequencing coverage of 12.5x would have been sufficient for SV detection in barley. When decreasing the sequencing coverage further, the precision and sensitivity to detect SV decreased considerably. Therefore, a sequencing coverage of 12.5x could be used to detect SV clusters in a small discovery panel as it was performed in our study. In a next step, a larger panel of hundreds of accessions could be used for genotyping the detected SV clusters based on a lower sequencing coverage. However, the performance of such two-step approaches needs first to be evaluated based on computer simulations.

The SV callers evaluated here were chosen based on former benchmarking studies in human (Cameron et al. 2019; Chaisson et al. 2019; Kosugi et al. 2019) as well as rice (Fuentes et al. 2019) and pear (Liu et al. 2020b). Across all SV types and SV length categories, we observed the highest precision and sensitivity for Manta and GRIDSS followed by Pindel with only marginally lower values (Table 2). This finding is in accordance with results of Cameron et al. (2019) for humans. In comparison with the results of Fuentes et al. (2019), we observed a considerably lower sensitivity and precision for Lumpy and NGSEP (Table 2). This difference in performance of the SV callers in rice and barley might be explained by the difference in genome length as well as the high proportion of repetitive elements in the barley genome (Mascher et al. 2017).

Despite the high sensitivity and precision observed for some SV callers, we observed even higher values when using them in combination (Table 2). This can be explained by the different detection principles such as paired-end reads, split reads, read depth, and local assembling that are underlying the different SV callers. Our observation indicates that a combined use of different short-read SV callers is highly

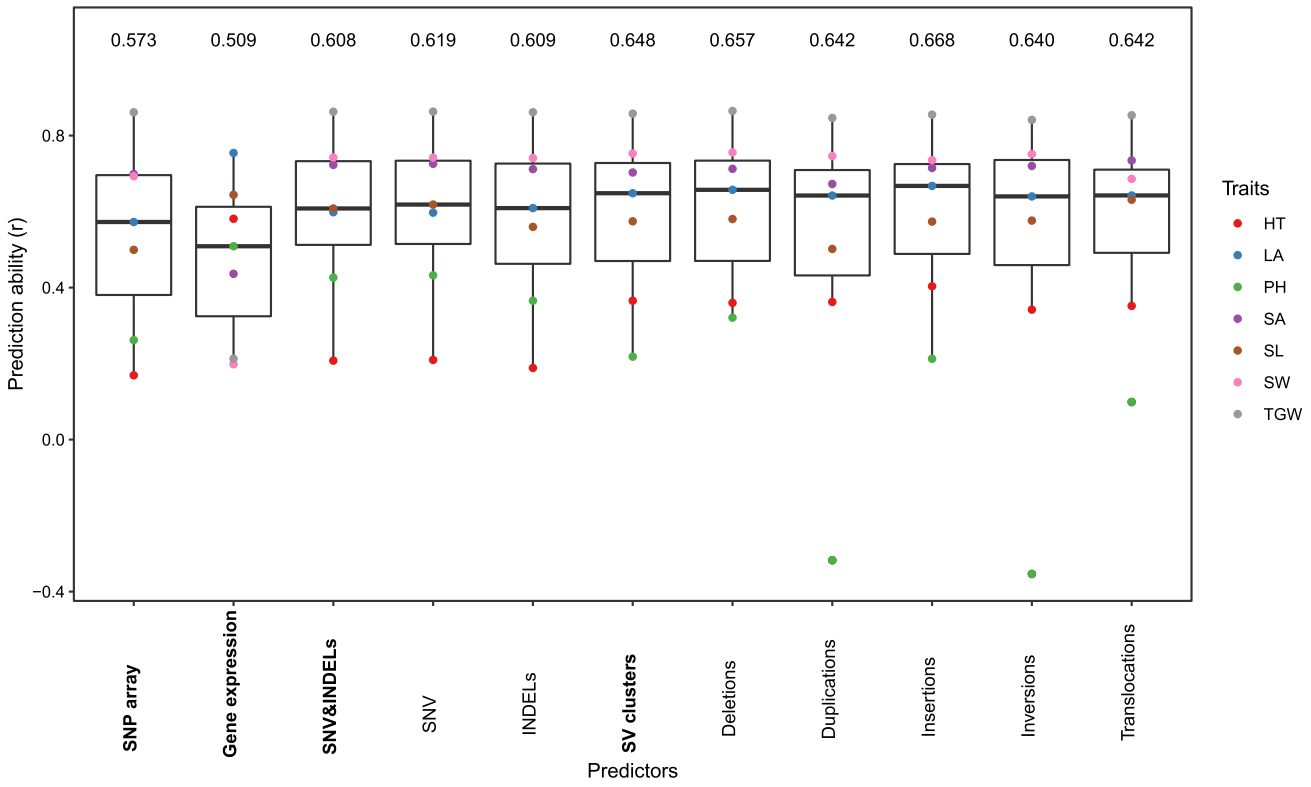


Fig. 6 Boxplot of the median prediction abilities across the seven traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), thousand grain weight (TGW) based on 23 inbreds using different predictors. The points in each box represent the medians of 200 five-fold cross-validation runs for each trait. The predictors were: features from SNP array, gene expression, single nucleotide variants (SNV) and small insertions and deletions (2–49 bp, INDELs), as well as structural variant (SV) clusters individually as well as combined together

recommended. This approach was then used for SV detection in the set of 23 spring barley inbreds.

Validation of SV in the barley genome

A PCR-based approach was used to validate a small subset of all detected SV. In accordance with earlier studies (Zhang et al. 2015; Yang et al. 2019; Guan et al. 2021), we evaluated the agreement between the detected SV and PCR results (Supplementary Fig. S3) for deletions and insertions up to 0.3 kb (Supplementary Fig. S4). For eleven out of the eleven SV, we observed a perfect correspondence.

Our PCR results further suggested that the SV callers were able to detect eight out of eleven deletions between 0.3 and 460 kb (Supplementary Fig. S5) based on the short-read sequencing of the non-reference parental inbred Unumli-Arpa. In four of the eleven PCR reactions, however, more than one band was observed. This was true three times for the non-reference genotype Unumli-Arpa and one time for Morex (Supplementary Fig. S5B). In two of the four cases, PCR indicated the presence of both SV states in one genome. This was true for Morex as well as Unumli-Arpa and might be due to the complexity of the barley genome which increases the potential for off-target amplification.

In conclusion, for 19 of the 22 tested SV (Supplementary Table S2), the SV detected in the non-reference parental inbred by the SV callers was also validated by PCR. This high validation rate implies in addition to the high precision and sensitivity values observed for SV detection in the computer simulations that the SV detected in the experimental data of the 23 barley inbreds can be interpreted.

Characteristics of SV clusters in the barley gene pool

Across the 23 spring barley inbreds that have been selected out of a world-wide diversity set to maximize phenotypic and genotypic diversity (Weisweiler et al. 2019), we have identified 458,671 SV clusters (Table 3). This corresponds to 1 SV cluster every 9,149 bp and corresponds to what was observed by Jayakodi et al. (2020). This number is in agreement with the number of SV clusters detected for cucumber (9,788 bp^{-1}) (Zhang et al. 2015) or peach (8,621 bp^{-1}) (Guan et al. 2021). Other studies have revealed a higher number of SV clusters than observed in our study. This might be due to the considerably higher number of re-sequenced accessions

Theoretical and Applied Genetics (2022) 135:3511–3529

in rice (214 bp^{-1}) (Fuentes et al. 2019), tomato (3,291 bp^{-1}) (Alonge et al. 2020), and grapevine (1,260 bp^{-1}) (Zhou et al. 2019).

The highest proportion of SV clusters detected in our study were deletions, followed in decreasing order by translocations, duplications, insertions, and inversions (Table 3). This is in disagreement with earlier studies where the frequency of duplications was considerably lower compared to that of insertions (Zhang et al. 2015; Zhou et al. 2019; Guan et al. 2021). Barley's high proportion of duplications compared to other crops may be due to its high extent of repetitive elements (Mascher et al. 2017).

In contrast to earlier studies in grapevine and peach (e.g., Zhou et al. 2019; Guan et al. 2021) we observed a strong non-uniform distribution of SV clusters across the genome. Only 14.5% of the SV clusters were located in pericentromeric regions, which make up 25.7% of the genome, whereas the rest was located distal of the pericentromeric regions (Fig. 2). This pattern was even more pronounced for SV hotspots, i.e., regions with a significantly ($P < 0.05$) higher amount of SV clusters than expected based on the average genome-wide distribution. Almost all SV hotspots (95.5%) were located distal of the pericentromeric regions (74.3% of the genome) where higher recombination rates are observed. Our observation indicates that the majority of SV clusters in barley might be caused by mutational mechanisms related to DNA recombination-, replication-, and/or repair-associated processes and might be only to a lower extent due to the activity of transposable elements. This is supported by the observation that, with the exception of translocations, only 1.4 to 25.2% of SV clusters were located in genome regions annotated as transposable elements (Table 3).

To complement our genome-wide analysis of barley SV clusters, we also examined their occurrence relative to genes and their association with gene expression.

Association of SV clusters with transcript abundance

About 60% of the SV clusters were detected in the intergenic space (Fig. 3). The remaining SV clusters were gene-associated and detected in regions either 5kb up- or downstream of genes (~30%) while ~10% were detected in introns and exons (Fig. 3). These values are in the range of those previously reported for rice (~75%, NA, exons: ~6%) (Fuentes et al. 2019), potato (~37%, ~37%, ~26%) (Freire et al. 2021), and peach (~52%, ~27%, ~21%) (Guan et al. 2021). The higher proportion of SV clusters in genic regions in potato and peach compared to the cereal genomes might suggest that SV clusters are more frequently associated with gene expression in clonally than in sexually propagated species. A possible explanation for this observation could be the degree of heterozygosity in clonal species, which is considerably higher compared to that in selfing species such as rice and barley. Hence, it is plausible that they better tolerate SV clusters close to genes.

Our study was based on 23 barley inbreds which confer a limited statistical power to detect SV cluster-gene expression associations. However, this leads not to an increased proportion of false positive associations. Therefore, the findings are discussed here.

We observed that the average absolute correlation coefficient of gene-associated SV clusters and global gene expression measured as loadings on the principal components was with 0.17 significantly ($P < 0.05$) different from 0 (Supplementary Fig. S12). In addition, 700 gene-associated SV clusters were individually associated ($P < 0.05$) with genome-wide gene expression. A further 1,976 alleles of gene-associated SV clusters were significantly ($P < 0.05$) associated with the expression of the corresponding 1,594 genes (Fig. 5). Additional support is given by the observation that despite SV clusters have a similar distribution across the genome as SNV, SV clusters covered more positions (in bp) of promoter regions than SNV (Fig. 4). These figures of significantly gene-associated SV clusters are in agreement with earlier figures for tomato (Alonge et al. 2020) and soybean (Liu et al. 2020a) and highlight the high potential of SV clusters to be associated with phenotypic traits.

Genomic prediction

Because of the limited number of inbreds included in this study, the power to identify causal links between SV clusters and phenotypes is low when considering only the 23 inbreds. However, instead of examining the association of individual SV clusters with phenotypic traits, we evaluated their potential to predict seven phenotypic traits in comparison with various other molecular features which is expected to provide reasonable information also with a limited sample size (Weisweiler et al. 2019).

We observed that the ability to predict these seven traits was higher for SV clusters compared to the benchmark data from a SNP array (Fig. 6). This might be explained by the considerably higher number of SV clusters than variants included in the SNP array. However, we observed the same trend when comparing the prediction ability of SV clusters to that of the much more abundant SNV & INDELs. This indicates that the SV clusters comprise genetic information that is not comprised by SNV & INDELs. Our result is supported by the observation that when examining the combination of SNV and INDELs with gene expression and SV clusters to predict phenotypic traits, an increase of the prediction ability was observed compared to the ability observed for the individual predictors (Supplementary Table S5). Furthermore, our observation of a different prediction ability

between SV clusters and SNV & INDELs can be explained by a lower extent of LD between SV clusters and linked SNV compared to that between SNV and linked SNV (Supplementary Table S4). These findings together illustrate the high potential of using SV clusters for the prediction of phenotypes in diverse germplasm sets. Such type of applications might be used also in commercial plant breeding programs. From a cost perspective such approaches will be realistic if SV detection is possible from low coverage sequencing. This might be possible when comprehensive reference sets of SV per species are available as was, for example, generated in our study for barley. However, this requires further research.

Usefulness of SV information for QTL fine mapping and cloning

The inbred lines included in our study are the parents of a new resource for joint linkage and association mapping in barley, the double round robin population (HvDRR, Casale et al. 2022). This population consists of 45 biparental segregating populations with a to total of about 4,000 recombinant inbred lines and is available from the authors upon reasonable request. The detailed characterization of the SV pattern of the parental inbreds, presented in this study, will therefore be an extremely valuable information for the ongoing and future QTL fine mapping and cloning projects exploiting one or multiple of the HvDRR sub-populations.

To illustrate this, we have mapped the naked grain phenotype in six HvDRR sub-populations (HvDRR03, HvDRR04, HvDRR20, HvDRR23, HvDRR44, HvDRR46) to chromosome 7H (7H:525,620,758-525,637,446). Taketa et al. (2008) discovered a 17 kb deletion harboring an ethylene response factor gene on chromosome 7H that caused naked caryopses in barley. In our study, two parental inbreds, namely Kharsila and IG128104, are naked barley. For both inbreds, the SV calls revealed the same 17 kb deletion on chromosome 7H. In contrast, the deletion was absent in the 21 other parental inbreds. This illustrates the potential of exploiting SV information of parental inbreds for gene QTL and gene cloning.

Furthermore, four indels which occur in the 5kb up-/downstream and genic regions of the VRS1 gene were significantly ($P < 0.01$) associated with the rowtype of the parental inbreds.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-022-04197-7.

Acknowledgements Computational infrastructure and support were provided by the Center for Information and Media Technology (ZIM) at Heinrich Heine University Düsseldorf.

Author Contribution Statement MW and BS designed and coordinated the project; TH extracted DNA and prepared the libraries; DVI contributed phenotypic data; MW, CA, and PW performed the analyses; MW and BS wrote the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111). The funders had no influence on study design, the collection, analysis and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication.

Data Availability Raw DNA sequencing data of the 23 barley inbreds have been deposited into the NCBI Sequence Read Archive (SRA) under the accession PRJNA77700. Raw mRNA sequencing data are available under the accession PRJNA534414. Data of gene expression, SNP array, adjusted entry means of phenotypes, INDELs, and SV clusters are available via figshare (https://doi.org/10.6084/m9.figshare.16802473). SNV data are available via zenodo (https://doi.org/10.5281/zenodo.6451025). Snakemake workflows are available via github (https://github.com/mw-qggp/SV_barley). Further scripts are available from the authors upon request.

Declarations

Competing interests The authors declare that they have no competing interests.

Ethics approval and consent to participate The authors declare that the experimental research on plants described in this paper complied with institutional and national guidelines.

Consent for publication All authors read and approved the final manuscript.

References

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR et al (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145-161.e23

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander

ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Baker M (2012) Structural variation: the genome's hidden architecture. Nat Methods 9:133–137

Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, Waugh R (2017) Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci 8:1792

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421

Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27:1–11

Cameron DL, Di Stefano L, Papenfuss AT (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun 10:3240

Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B (2022) Genomic prediction of the recombination rate variation in barley—a route to highly recombinogenic genotypes. Plant Biotechnol J 20:676–690

Chaisson MJ, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D et al (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10:1784

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM (2017) The impact of structural variation on human gene expression. Nat Genet 49:692–699

Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Yuan Wang Z, Wang A, Wang X, Wang J, Wei MH, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Lai Cheng M, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Ni Tint N, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Deslattes Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Craig-Holmes AP, Moore FB, Shaw MW (1973) Polymorphism of human C-band heterochromatin. I. Frequency of variants. Am J Hum Genet 25:181–192

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN (2021) How the pan-genome is changing crop genomics and improvement. Genome Biol 22:3

Duitama J, Quintero JC, Cruz DF, Quintero C, Hubmann G, Foulquié-Moreno MR, Verstrepen KJ, Thevelein JM, Tohme J (2014) An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. Nucleic Acids Res 42:e44

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, Höps W, Ashraf H et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372:eabf7117

Elyanow R, Wu HT, Raphael BJ (2018) Identifying structural variants using linked-read sequencing data. Bioinformatics 34:353–360

Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 Genes|Genomes|Genetics 211:1405

Freire R, Weisweiler M, Guerreiro R, Baig N, Hüttel B, Obeng-Hinneh E, Renner J, Hartje S, Muders K, Truberg B, Rosen A, Prigge V, Bruckmüller J, Lübeck J, Stich B (2021) Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. G3 Genes|Genomes|Genetics 11:jkab330

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, Mauleon R, Alexandrov N (2019) Structural variants in 3000 rice genomes. Genome Res 29:870–880

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546–567

Gong T, Hayes VM, Chan EK (2021) Detection of somatic structural variants from short-read next-generation sequencing data. Brief bioinform 22:1–15

Gouesnard B (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, Zhao J, Jiang Q (2021) Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape. Genome Biology 22:13

Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH (2010) Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breed 129:271–279

Hill WG, Robertson A (1968) Linkage disequilibrium among neutral genes in finite populations. Theor Appl Genet 38:226–231

Jacobs PA, Strong JA (1959) A case of human intersexuality having a possible XXY sex-determining mechanism. Nature 183:302–303

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang XQ, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Fiebig A, Budak H, Xu D et al (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289

Karaoğlanoğlu F, Ricketts C, Ebren E, Rasekh ME, Hajirasouliha I, Alkan C (2020) VALOR2: characterization of large-scale structural variants using linked-reads. Genome Biol 21:72

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S (2021) Sustainable data analysis with Snakemake. F1000Research 10:33

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 20:117

Kou Y, Liao Y, Toivainen T, Lv Y, Tian X, Emerson JJ, Gaut BS, Zhou Y (2020) Evolutionary genomics of structural variation in asian rice (Oryza sativa) domestication. Mol Biol Evol 37:3507–3524

Kühl MA, Stich B, Ries DC (2021) Mutation-simulator: fine-grained simulation of random mutations in any genome. Bioinformatics 37:568–569

Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, Jia J, Kong X (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol 196:282–291

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020) Pan-genome of wild and cultivated soybeans. Cell 182:162–176

Liu Y, Zhang M, Sun J, Chang W, Sun M, Zhang S, Wu J (2020) Comparison of multiple algorithms to reliably detect structural variants in pears. BMC Genomics 21:61

Luo R, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC (2017) LRSim: a linked-reads simulator generating insights for better genome partitioning. Comput Struct Biotechnol J 15:478–484

Mahmoud M, Gobet N, Cruz-dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol 20:246

Manolov G, Manolov Y (1972) Marker band in one chromosome 14 from Burkitt lymphomas. Nature 237:33–34

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, Fiddes IT, Galvin B, Heaton H, Herschleb J, Hindson C, Holt E, Jabara CB, Jett S, Keivanfar N, Kyriazopoulou-Panagiotopoulou S, Lek M et al (2019) Resolving the full spectrum of human genome variation using linked-reads. Genome Res 29:635–645

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, Magalhaes JV, Piñeros MA, Schatz MC, Wing RA, Kochian LV (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A 110:5241–5246

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, Mccooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, Gundlach H, Boston LB, Tulpová Z, Holden S, Hernández-Pinzón I, Scholz U, Mayer KF, Spannagl M, Pozniak CJ, Sharpe AG, Simková H, Moscou MJ, Grimwood J, Schmutz J, Stein N (2021) Long-read sequence assembly: a technical evaluation in barley. Plant Cell 33:1888–1906

McColgan P, Tabrizi SJ (2018) Huntington's disease: a clinical review. Eur J Neurol 25:24–34

Mitelman F, Catovsky D, Manolova Y (1979) Reciprocal 8;14 translocation in EBV-negative B-cell acute lymphocytic leukemia with Burkitt-type cells. Int J Cancer 24:27–33

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KF, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20:284

Morisse P, Legeai F, Lemaitre C (2021) LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. bioRxiv. https://doi.org/10.1101/2021.03.25.437002

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KFX, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14:R58

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, DeJong PJ, Yorke JA, Salzberg SL, Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5' upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. Mol Breed 31:27–37

Nowell P, Hungerford D (1960) Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 25:85–109

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:333–339

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208:1373–1385

Schüle B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, Liu M, Byrne C, Gopi R, Huang N, Langston JW, Clark T, Gil FJJ, Ashizawa T (2017) Parkinson's disease associated with pure ATXN10 repeat expansion. npj Parkinson's Dis 3:27

Stich B, Möhring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 1783:1745–1754

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJ, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HY, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318:1446–1449

Taketa S, Amano S, Tsujino Y, Sato T, Saisho D, Kakeda K, Nomura M, Suzuki T, Matsumoto T, Sato K, Kanamori H, Kawasaki S, Takeda K (2008) Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. Proc Natl Acad Sci U S A 105:4062–4067

The Arabidopsis Genome Iniative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 48:796–815

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3-new capabilities and interfaces. Nucleic Acids Res 40:e115

VanRaden P (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann JC, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Wang O, Chin R, Cheng X, Yan Wu MK, Mao Q, Tang J, Sun Y, Anderson E, Lam HK, Chen D, Zhou Y, Wang L, Fan F, Zou Y, Xie Y, Zhang RY, Drmanac S, Nguyen D, Xu C, Villarosa C, Gablenz S, Barua N, Nguyen S, Tian W, Liu JS, Wang J, Liu X, Qi X, Chen A, Wang H, Dong Y, Zhang W, Alexeev A, Yang H, Wang J, Kristiansen K, Xu X, Drmanac R, Peters BA (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res 29:798–808

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. Genome Res 27:757–767

Weisweiler M, Montaigu AD, Ries D, Pfeifer M, Stich B (2019) Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue RNA sequencing and their power to predict phenotypic traits. BMC Genomics 20:787

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, Wang Y, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Li W, Liu L, Jackson D, Yan J (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51:1052–1059

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, Sun H, Jiao C, Blakely R, Pan J, Cai R, Luo R, Van de Peer Y, Jacobsen E, Fei Z, Huang S (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27:1595–1604

Zheng X, Medsker B, Forno E, Simhan H, Juan C, Sciences R (2016) Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. Nat Biotechnol 34:303–311

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS (2019) The population genetics of structural variants in grapevine domestication. Nat Plants 5:965–979

4 Affordable, accurate and unbiased RNA sequencing by manual library miniaturization: A case study in barley

This manuscript was published in Plant Biotechnology Journal in July 2023. Supplementary material is available online.

Authors:
Christopher Arlt, Thorsten Wachtmeister, Karl Köhrer, and Benjamin Stich.

Own contribution: First author. I conducted the experiment, performed data analysis, and wrote the manuscript.

Plant Biotechnology Journal (2023) 21, pp. 2241–2253 doi: 10.1111/pbi.14126

Affordable, accurate and unbiased RNA sequencing by manual library miniaturization: A case study in barley

Christopher Arlt[1], Thorsten Wachtmeister[2], Karl Köhrer[2] and Benjamin Stich[1,3,4,*,a]

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[2]Genomics & Transcriptomics Laboratory, Biological and Medical Research Centre (BMFZ), Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[3]Cluster of Excellence on Plant Sciences (CEPLAS), Duesseldorf, Germany
[4]Max Planck Institute for Plant Breeding Research, Cologne, Germany

Received 16 November 2022;
revised 12 May 2023;
accepted 1 July 2023.
**Correspondence* (Tel: +49 38209 45201; Fax: +49 38209 45222; email Benjamin.Stich@julius-kuehn.de)
[a]Present address: Institute for Breeding Research on Agricultural Crops, Julius Kühn Institute (JKI) - Federal Research Centre for Cultivated Plants, Sanitz, Germany

Keywords: methods and techniques, RNA sequencing, library preparation, miniaturization, plant breeding, plant genetics.

Summary

We present an easy-to-reproduce manual miniaturized full-length RNA sequencing (RNAseq) library preparation workflow that does not require the upfront investment in expensive lab equipment or long setup times. With minimal adjustments to an established commercial protocol, we were able to manually miniaturize the RNAseq library preparation by a factor of up to 1:8. This led to cost savings for miniaturized library preparation of up to 86.1% compared to the gold standard. The resulting data were the basis of a rigorous quality control analysis that inspected: sequencing quality metrics, gene body coverage, raw read duplications, alignment statistics, read pair duplications, detected transcripts and sequence variants. We also included a deep dive data analysis identifying rRNA contamination and suggested ways to circumvent these. In the end, we could not find any indication of biases or inaccuracies caused by the RNAseq library miniaturization. The variance in detected transcripts was minimal and not influenced by the miniaturization level. Our results suggest that the workflow is highly reproducible and the sequence data suitable for downstream analyses such as differential gene expression analysis or variant calling.

Introduction

Next-generation sequencing (NGS) technologies have been evolving rapidly in the last two decades and continue to do so (Mardis, 2011; McCombie *et al.*, 2019). NGS can be used for a variety of different applications and is nowadays an integral part of many genetic research projects. This became possible in part by steadily decreasing sequencing costs (Wetterstrand, 2021).

This development not only enhanced the possibilities of whole genome sequencing (Auton *et al.*, 2015; Chung *et al.*, 2017; Harris and Wailan Alexander, 2021; Linderman *et al.*, 2016) but also mRNA sequencing projects (Li, 2021; Stark *et al.*, 2019). Because the transcriptome size is relatively consistent between species, the sequencing costs for species with large genomes benefit greatly by focusing on the protein-coding part of the genome. Additionally, targeting only the mRNA for sequencing is a useful complexity reduction when investigating the genotype–phenotype relationship (Jehl *et al.*, 2021; Piskol *et al.*, 2013; Shomroni *et al.*, 2022; Wang *et al.*, 2021). The relatively low complexity of mRNA libraries and the increased read output of large sequencing platforms expanded the multiplexing potential in RNA sequencing (RNAseq) projects allowing for 384+ samples to be pooled and sequenced in the same sequencing reaction. This results in a cost distribution shift, making the library preparation step the most expensive part of many RNAseq projects. The pressure to reduce the costs of this step is therefore rising and many approaches have been developed to do exactly that (Alpern *et al.*, 2019; Bagnoli *et al.*, 2018; Foley *et al.*, 2019; Hashimshony *et al.*, 2016; Hou *et al.*, 2015; Islam *et al.*, 2012; Kumar *et al.*, 2012; Pallares *et al.*, 2019; Picelli *et al.*, 2013; Shishkin *et al.*, 2015). One way to save costs during the preparation of RNAseq libraries is to switch from commercial protocols to previously published custom protocols. The latter often implement novel techniques to optimize the procedure and save costs. If these approaches are successful enough, commercial adaptations are developed, as was the case with the examples mentioned below.

The traditional protocols create RNAseq libraries using full-length mRNA molecules (Hou *et al.*, 2015; Islam *et al.*, 2012; Kumar *et al.*, 2012; Picelli *et al.*, 2013; Shishkin *et al.*, 2015). A more cost-efficient alternative is to create libraries of the 3′ or 5′ end of the mRNA exclusively (Foley *et al.*, 2019; Macosko *et al.*, 2015; Pallares *et al.*, 2019; Vahrenkamp *et al.*, 2019). Some protocols employ early multiplexing to further reduce hands-on time and costs within the prime end enriched library preparation methods (Alpern *et al.*, 2019; Bagnoli *et al.*, 2018; Hashimshony *et al.*, 2016; Soumillon *et al.*, 2014). In most full-length mRNA protocols, the multiplex bar code is part of the adapter sequence which is added to the library fragments late in the library preparation workflow and the sample pooling is conducted after amplification and clean-up. When utilizing early multiplexing, a unique bar code is added to the sequences in one of the initial steps of the protocol. This enables early multiplexing and reduces the number of samples handled during the remaining library preparation steps. While both strategies are a good way to reduce costs, they limit the application of the resulting data for further analyses for example, genomic variant calling or novel transcript identification. Additionally, early multiplexing strategies make it impossible to re-sequence individual samples.

A different approach is most commonly known as miniaturization. It involves the reduction of the utilized reagent volume during the library preparation using commercial protocols. Most

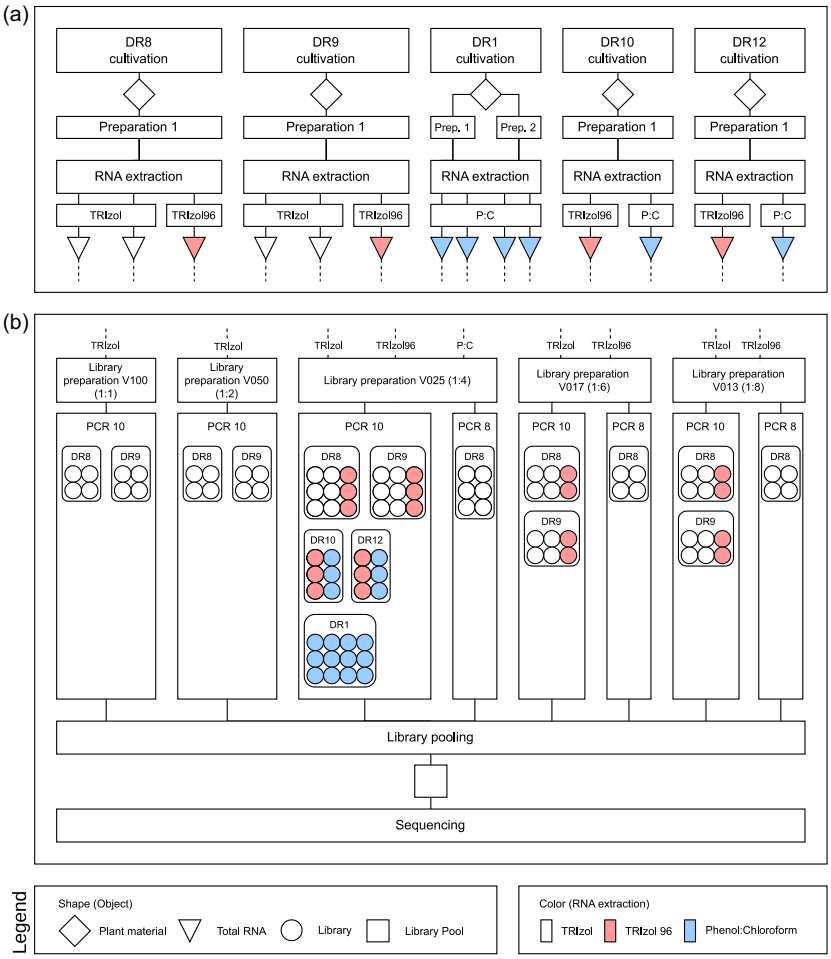


Figure 1 Overview of experimental design and miniaturization levels of the 96 samples. (a) Five different recombinant inbred lines (RIL) were cultivated and harvested (diamond). For all RILs except DR1, a single plant material preparation was used for one or more total RNA extractions using different methods (white: TRIzol, red: TRIzol-96, blue: Phenol:Chrloroform (P:C)), resulting in a total of 14 different total RNA samples (triangle). The DR1 plants were used to create coarse (Prep. 1) and fine (Prep. 2) ground plant material. (b) We tested the library preparation miniaturization levels 1:1 (V100, original), 1:2 (V050), 1:4 (V025), 1:6 (V017) and 1:8 (V013). The PCR cycles were reduced from 10 (PCR 10) to 8 (PCR 8) for a subset of the samples. For each RIL in each miniaturization and number of PCR cycles, 2–3 RNA extraction and library preparation replicates were created. In total, 96 RNA sequencing libraries were created (circles) and combined to a single library pool (square).

described workflows use full-length mRNA protocols combined with liquid handling automatization (e.g. Jaeger *et al.*, 2020; Mayday *et al.*, 2019; Mildrum *et al.*, 2020; Mora-Castilla *et al.*, 2016). Adding automatization to the miniaturization workflow has two major advantages. First, all automated preparation steps reduce hands-on time and therewith labour costs. Second, the inherent reduction in sample-to-sample variance by replacing the less error-prone hands-on steps increases the level of accuracy and precision (Tegally *et al.*, 2020). However, the investment costs of acquiring all the lab equipment required for an automated library preparation workflow are high and, thus, not feasible for many research groups. To our knowledge, with the exception of Li *et al.* (2019), who miniaturized the DNA library preparation of *E. coli* genomes, no studies are available on the capabilities of manual miniaturization.

While it is possible to reduce the costs of the library preparation step in many different ways as was outlined above, it is crucial that the quality of the resulting data sets is not impaired and has no negative impact on downstream analyses (Aigrain *et al.*, 2016; Alberti *et al.*, 2014; Dabney and Meyer, 2012; McNulty *et al.*, 2020; Romero *et al.*, 2014). However, to the best of our knowledge, no comprehensive characterization of library complexity and biases was performed for manual library preparation miniaturizations.

Here we present an easy-to-reproduce, manual miniaturized full-length mRNA sequencing library preparation workflow that does not require the upfront investment in expensive lab equipment. A miniaturization level of up to 1:8 was tested which reduced the library preparation costs significantly. In addition, we provide the results of a wide set of quality control analyses, evaluating the impact of the miniaturization on the resulting sequencing data.

Results

To evaluate the success of the library preparation miniaturization workflow, 96 samples were prepared without miniaturization (1:1, V100) and the miniaturization levels 1:2 using 50% of all reagents (V050), 1:4 using 25% of all reagents (V025), 1:6 using 17% of all reagents (V017) and 1:8 using 13% of all reagents (V013) (Figure 1). Five different genotypes from three different recombinant inbred line (RIL) populations were used to evaluate the miniaturization workflow. Two of the five RIL (DR8 and DR9) were included in all miniaturization levels to allow orthogonal comparisons. DR10 and DR12 were prepared using different RNA extraction methods and DR1 was used to analyse the potential impact of plant material coarseness on the workflow. The sequencing results were analysed with regards to library quality and its properties in common downstream analyses.

RNA extraction and library pool

All total RNA samples except the TRIzol-96 RNA extractions were evaluated using the Fragment Analyzer. The average RNA quality

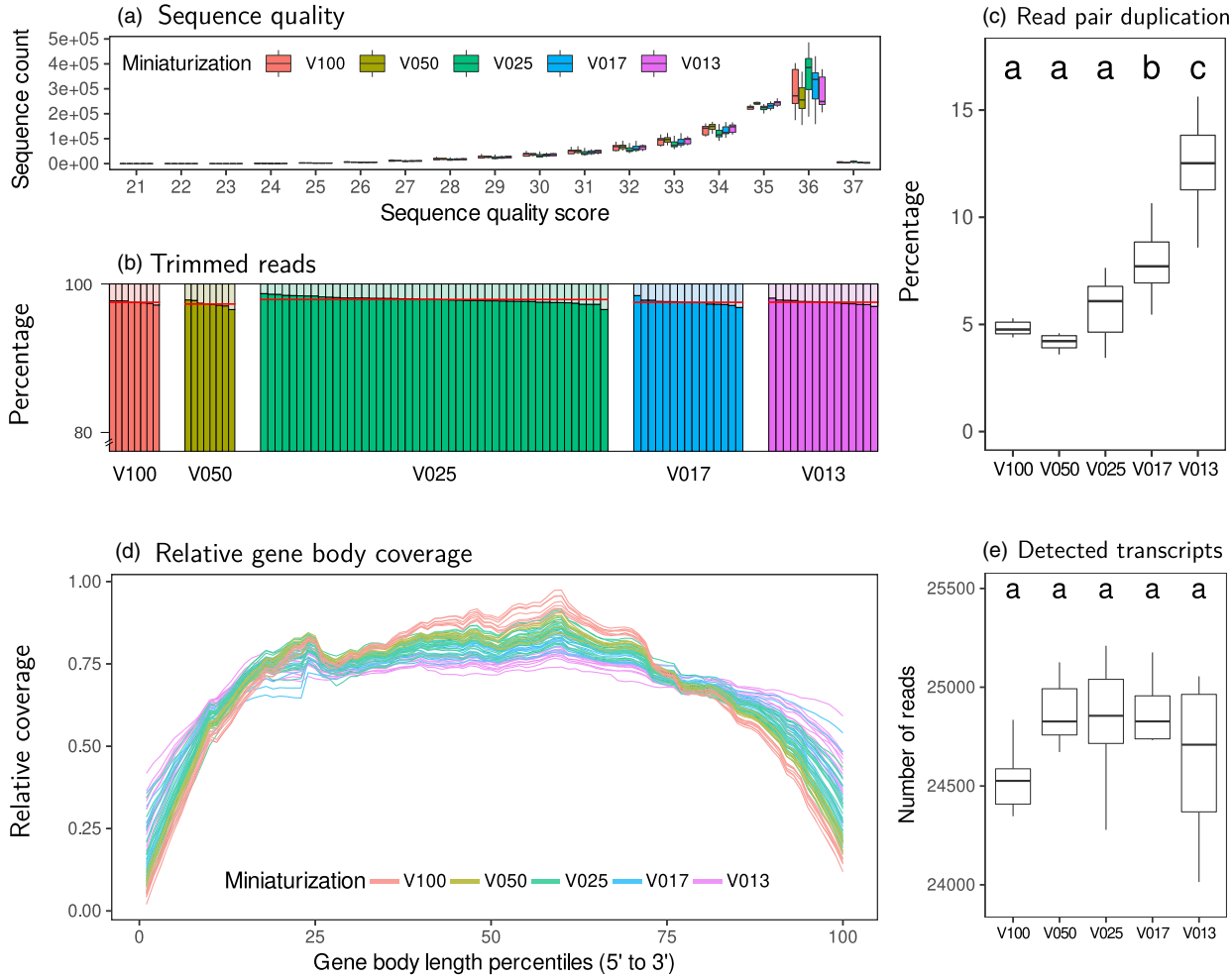


Figure 2 Overview of sequencing and library quality metrics. (a) Mean per base quality score for all 94 samples coloured by miniaturization level. (b) Percentage of reads after trimming (dark colour) per miniaturization level. The mean rate of remaining reads per miniaturization level shown as red line. (c) Total percentage of read pair duplicates by miniaturization. Miniaturization marked by the same letter are not significantly ($\alpha = 0.05$) different from each other. (d) Gene body coverage for the highly expressed genes (>90 transcript expression quantile) shown as percentage of reads located in each of the 100 gene segments starting at the 5′ end across all 94 samples coloured by miniaturization. (e) Number of unique transcripts detected by miniaturization.

number (RQN) score for the Phenol:Chloroform extraction method was the lowest with an average of 5.1 (coefficient of variation (CV) 0.13). The TRIzol extractions have an average RQN score of 7.4 (CV 0.04) (Figure S1). We also used the Fragment Analyzer to characterize the size distribution of the final library pool that was sequenced. While the fragments were with a peak at 392 bp smaller than aimed for, the size distribution itself was as expected (Figure S2).

Sequencing and library quality assessment

To evaluate the impact of the miniaturization on the sequencing process itself, we compared the mean per sequence quality score of all reads and did not find any difference between samples and miniaturizations in that regard (Figure 2a). The same was true for the per base *n* count, sequence length distribution, the per base sequence content and the adapter content (Figure S3). In addition, no negative trend in the trimming rates was observed between miniaturizations (Figure 2b). The mean read pair duplication rate was 7.1%. The highest read pair duplication ratio and significantly ($P < 0.001$) different from the remaining miniaturization levels in both RIL was observed for V013 (12.9%) (Figure 2c). For one of the two RIL, the miniaturization level V017 was significantly ($P < 0.001$) different from the rest. We assessed whether a general 5′ gene body coverage bias existed in our data set, but could not find one. The distribution did also not differ between miniaturizations (Figure 2d). Random RNA fragmentation was tested by comparing the nucleotide compositions around the fragmentation site for all miniaturization levels (Figure S4). Significant ($P < 0.001$) differences between un-miniaturized samples and miniaturized samples were observed. However, these were not consistent between RILs DR8 and DR9 with the exception of eight positions comparing the miniaturization levels V100 and V013.

The number of transcripts detected per sample ranged between 24 500 and 25 000 and did not significantly ($P > 0.05$) differ between miniaturizations (Figure 2e) when comparing random sub samples with 2 million reads. Pearson correlation coefficients of the read counts between pairs of miniaturization levels were calculated for DR8 and DR9. For both RIL, the highest similarity was observed between the V025 and

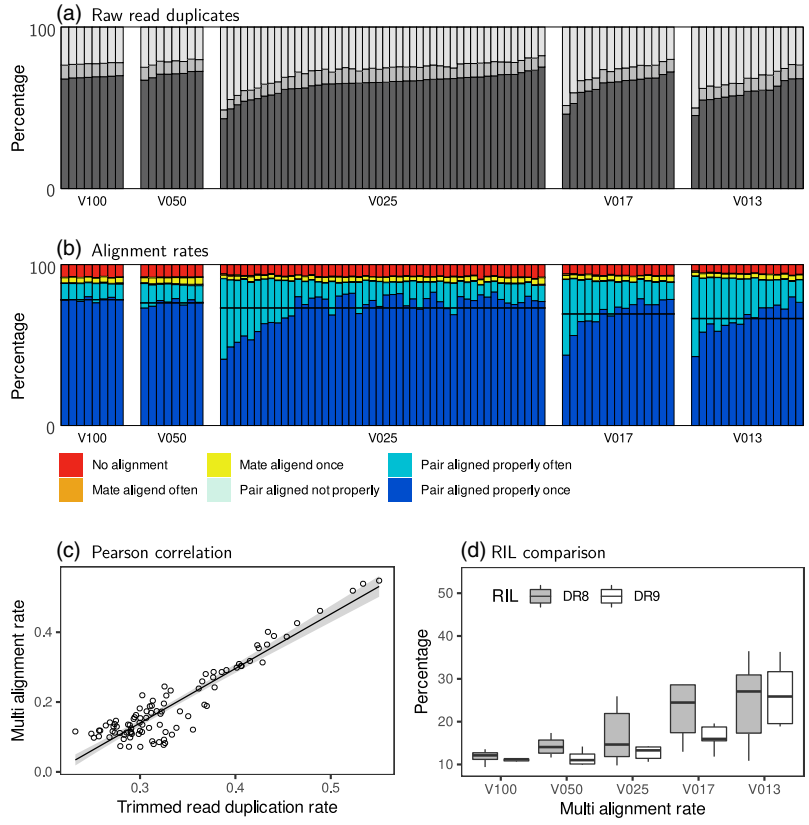


Figure 3 Correlation between read duplication rate and multi-alignment rate. (a) Trimmed read duplication rate of a subset with 1 million reads of all 94 samples coloured by uniqueness and grouped by miniaturization (Unique reads: bottom, dark grey; unique duplicated reads: middle, grey; remaining duplicated reads: top, light grey). (b) Alignment statistics for each sample grouped by miniaturization and coloured by alignment type. (c) Pearson correlation between multi-alignment rate and trimmed read duplication rate. (d) Comparison of multi-alignment rate between recombinant inbread lines (RIL) DR8 and DR9 for each miniaturization level.

V050 samples (DR8: $r = 0.9977$; DR9: $r = 0.9964$) (Table S1). The least similar sample groups were V100-V013 for DR8 ($r = 0.9613$) and V100-V017 for DR9 ($r = 0.9655$). We detected a high similarity between the read counts of replicates. The Pearson correlation coefficients for DR8 and DR9 were calculated for each miniaturization and replication type separately. The library replicates ranged from 0.999 to 0.977 and for RNA extraction replicates from 0.999 to 0.991 (Table S2).

While we focused on evaluating the impact of the miniaturization on the libraries, we also examined the impact of the number of PCR cycles, RNA extraction method and degree of plant material grinding. Therefore, we looked for differences in the rate of duplicated read pairs and the number of detected transcripts in these categories. Two of the six tests resulted in a significant ($P < 0.05$) difference in at least one group. First, the custom phenol:chloroform (HTP96) extraction method resulted in a significantly ($P < 0.001$) lower number of detected transcripts compared to the TRIzol method (0.77%). Secondly, the coarse grinding of plant material resulted in a significantly ($P < 0.001$) lower number of detected transcripts compared to the fine grinding of plant material (0.81%).

Increased variability in highly miniaturized samples

When comparing the rate of raw read duplications, we observed an increased variability between the samples within miniaturization levels above 1:4. Without miniaturization (V100) or a minimal miniaturization level (V050), the number of unique reads varied between 67.1% and 72.6% (Table S3). With higher miniaturization levels (V025, V017, V013) the variability between samples increased. For these samples, the rate of unique reads was between 45.2% and 75.3% (Figure 3a). The average rate of unique reads for DR8 and DR9 dropped from 69% (±0.68%) at V100 to 59% (±5.70%) for V013. Furthermore, the rate of uniquely aligned reads was highest for V100 (78% ± 1.22%) and lowest for V013 (66% ± 9.29%) and did therefore show the same trend as the rate of raw read duplications (Figure 3b). The overall alignment rate increased slightly with increasing miniaturization levels from V100 (92% ± 0.31%) to V013 (95% ±0.63%). The proportion of reads that were not properly aligned (once or multiple times) was similar across all miniaturizations (Table S4). The correlation coefficient between the raw read duplicates and read pair duplicates ($r = 0.52$, $P < 4.46e-08$) was lower than the correlation coefficient between raw read duplication rate and multi-alignment rate ($r = 0.93$, $P < 2.2e-16$) (Figure 3c). We also observed an impact of the genotype on the rate of multi-aligned reads (Figure 3d). In all miniaturizations, RIL DR9 had lower rates than DR8 but the differences were not significant ($P > 0.05$). The proportion of duplicated sequences with more than 10 identical duplicates was considerably increased for V025, V017 and V013 compared to V100 and V050 (Figure S5a). Additionally, for V017 and V013 the proportion of duplicated reads with more than 100 identical duplicates was further increased compared to the other miniaturizations.

Characterizing the multi-aligned reads

To further investigate the increased between-sample variability that was detected in V025, V017 and V013 miniaturizations, we created two data subsets. First, we subsetted the alignment results including only reads that were mapped multiple times during the alignment. Those read IDs were used to create the

subset of multi-aligned reads. Comparing the rate of raw read duplications between the subset of multi-aligned reads and the total data set showed that the mean duplication rate increased by 36% in the subset (Figure S6a). For a considerable portion of these duplicates, more than 10 reads of the same sequence were present (Figure S5b). The mean proportion of all duplicated reads included in the subset of multi-aligned reads was around 40%, compared to only 7% of all unique reads (Figure S6b).

The origin of the multi-aligned reads was analysed by investigating gene annotation, transposable elements (TEs) sequence overlap and rRNA contamination. A GO term enrichment analysis between the subset of multi-aligned reads and the total data set resulted in multiple significantly underrepresented genes related to TE activity (biological process: 'RNA-dependent DNA biosynthetic process'; molecular function: 'RNA-directed DNA polymerase activity', 'RNA–DNA hybrid ribonuclease activity'). The most overrepresented genes were related to photosynthesis (various biosynthetic process, cellular compartment and molecular function annotations) and transcription (biological process: 'regulation of transcription by RNA polymerase II'; cellular compartment: 'mediator complex') (Figure S7).

On average, more than 55% of the multi-aligned reads could not be assigned to an annotated transcript (no feature, NF). This was significantly higher ($P < 0.001$) than the 16% of reads in the total data set (Figure S8a). The rate of NF reads in the subset of multi-aligned reads did not correlate well with the multi-alignment rate (Figure S8b). While the proportion of NF reads that were multi-aligned varied (82.4%–20.2%), the number of uniquely aligned NF reads remained constant between all 94 samples (4.4% ± 1.0%) (Figure S8c).

The rate of TE reads between the subset of multi-aligned reads and the total data set was significantly increased ($P < 0.001$) (Figure S9a). The rate of TE reads was positively correlated with the multi-alignment rate (Figure S9b). Nevertheless, on average 17% of TE reads were not multi-aligned (Figure S9c). The variability between samples was highest for the TE reads that aligned multiple times (9.6% ± 7.1%).

Lastly, two different rRNA reference sequence libraries were created and the subset of multi-aligned reads was aligned against them. The *Hordeum vulgare* rRNA reference library showed the highest overall alignment rate with most read pairs aligning multiple times against the reference sequences (Figure S6c,d). For both rRNA reference sequence libraries, the subset of multi-aligned reads showed a significantly higher alignment rate ($P < 0.001$) than the total data set (Figure 4a). The Pearson correlation coefficient between the rRNA alignment rate and the multi-alignment rate was 0.999 for the subset of multi-aligned reads and 0.996 for the total data set (Figure 4b). While the proportion of multi-aligned reads that were of rRNA origin varied (93.9%–16.5%), the number of multi-aligned non-rRNA reads remained constant between all 94 samples (3.6% ± 0.5%) (Figure 4c). Additionally, only a small proportion of rRNA reads were uniquely aligned (1.9% ± 2.2%).

Variant calling and differential expression analyses

While the miniaturization did not considerably change the overall number of detected SNPs, the ratio between reference SNPs and alternative SNPs changed in many but not all miniaturization scenarios in favour of an increase in alternative SNPs in higher miniaturizations. However, the change was not consistent and no clear trend was observed when increasing the miniaturization level (Table S5).

We used a principal component analysis (PCA) to examine the data set's capability to cluster the samples based on genetic differences. When using read count data, the first two principal components explained 33.7% of the variance (Figure 5a). When using the SNP data set, the first two principal components explained 58.6% of the variance (Figure 5b). Based on both data sets, we could show that all samples of the same population clustered together. Additionally, the SNP data differentiated each of the five RILs. The samples which were prepared using the same miniaturization level did not cluster together across RIL (Figure S10).

The mean proportion of detected transcripts between miniaturizations for DR8 and DR9 in a 2 million read subset were very similar (30%–31%). V100 had the lowest proportion of detected transcripts with 30.0% which was significantly ($P = 0.042$) less then V050 (30.5%). We created lists of consensus transcripts separately for each miniaturization of DR8 and DR9. The resulting number of detected transcripts within replicates of RIL DR8 and DR9 for each miniaturization level varied between 15 428 (V100, DR9) and 16 930 (V025, DR8). Lastly, we characterized the overlap between the transcripts of each group resulting in 81.2% and 80.8% of all detected transcripts present in all miniaturizations for DR8 and DR9, respectively (Figure 6). For both RIL the second biggest group of transcripts was detected in all miniaturization levels except V100. Only 159 (DR8) and 143 (DR9) transcripts were exclusively present in V100.

Discussion

The workflow we describe in this manuscript reduces RNA library preparation costs by miniaturizing the process by up to 1:8 of the original reagent volume of the commercial kit. However, in order to ensure that the proposed modifications to the commercial protocol did not decrease the quality of the resulting library, an in-depth characterization was required. While many automated miniaturizations have been shown to have negligible impact on the library quality (Jaeger *et al*., 2020; Kong *et al*., 2019; Mildrum *et al*., 2020; Mora-Castilla *et al*., 2016), we used a procedure without automation. Therefore one has to investigate the practicality of the workflow considering the potentially increased pipetting error variance and potentially decreased reproducibility by manually handling <2 μL volumes.

Quality control: Library complexity and biases

In a first step, we had to make sure that our modifications to the protocol had no negative impact on the sequencing process. The per sequence quality scores, per base *n* content and sequencing length distribution of V100 did not significantly ($P > 0.05$) out-perform miniaturized samples (Figures 2a and S3). Additionally, the proportion of reads that were discarded by trimming was comparable between all samples regardless of the miniaturization level (Figure 2b). These observations led to the conclusion that the miniaturization did not have a negative impact on the sequencing process itself.

Next, we investigated additional quality metrics that characterize the library properties directly. Particularly, the library complexity and the potential library biases were investigated. Library complexity, or the number of unique molecules in solution, represents the potential of the given library to produce a complete picture of the genome or transcriptome and unravels potential problems during the library preparation, for example, if a significant number of unique molecules was lost. One way to

 21, 2241–2253

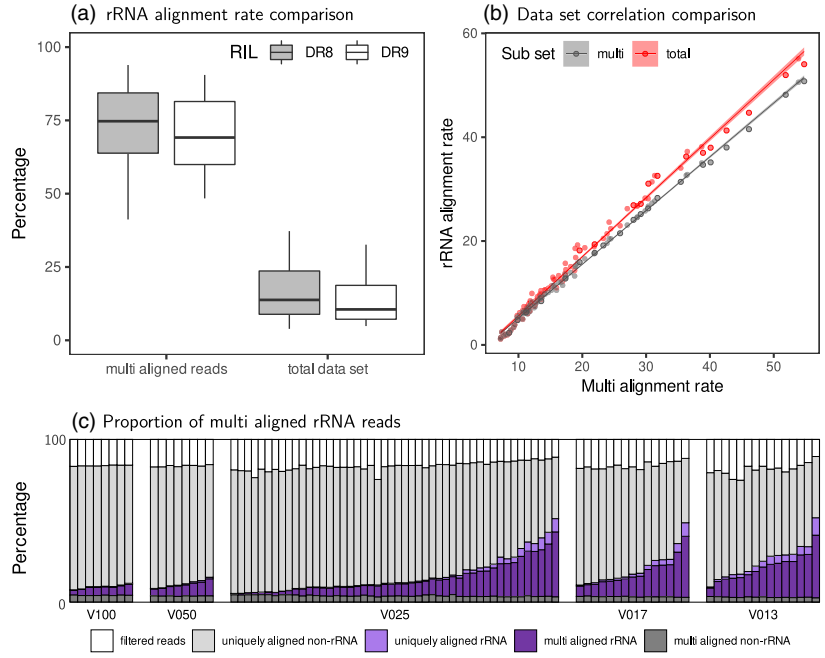


Figure 4 Investigation of the rRNA origin in the subset of multi-aligned reads. (a) Comparison of rRNA alignment rates using the *Horedum vulgare* rRNA reference library between the subset of multi-aligned reads and the total data set (recombinant inbred line (RIL) DR8 in grey and DR9 in white). (b) The correlation between the multi-alignment rate and the rRNA alignment rate for the subset of multi-aligned reads (multi, grey) and the total data set (total, red). (c) The proportion of filtered reads (white), uniquely aligned non-rRNA reads (light grey)/rRNA reads (light purple) and multi-aligned rRNA reads (dark purple)/non-rRNA reads (dark grey).

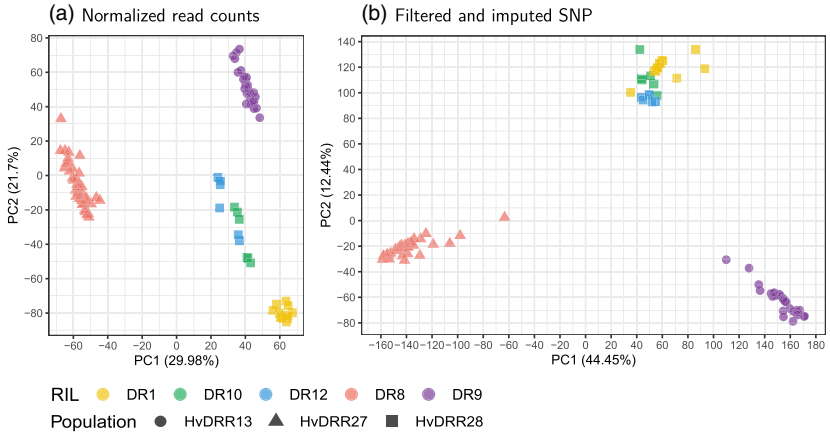


Figure 5 Principal component analysis based on RNA sequencing data of the five recombinant inbred lines (RIL). (a) Normalized read counts used as a basis. (b) Filtered and imputed SNPs used as a basis. PC 1 and PC 2 are the first and second principal component, respectively, and the number in parentheses refers to the proportion of variance explained by the principal components in percent.

assess library complexity is to compare read pair duplication rates (Alberti *et al*., 2014). A reduced number of unique molecules statistically increases the number of PCR duplicates included in the sequencing pool and therefore increases the probability for read pair duplication in the resulting sequencing data and at the same time reduces the number of unique transcripts that can be discovered. Alternatively, when analysing RNA sequencing data, the number of detected genes can be evaluated (Mereu *et al*., 2020). In our results, the mean read pair duplication was 7.1% with only the V013 miniaturizations showing significantly ($P < 0.001$) increased duplication rates for both RIL (Figure 2c). Therefore, V025 is an attractive miniaturization factor which makes most efficient use of the sequencing while at the same time resulting in a major cost decrease in the library preparation costs. Nevertheless, the overall read pair duplication rate observed for all miniaturization levels was comparable to those of previous studies (Bansal, 2017; Fu *et al*., 2018; Parekh *et al*., 2016). The number of unique detected transcripts did not differ significantly ($P > 0.05$) between all miniaturization levels (Figure 2e) and both V017 and V013 did not separate in our PCA using read count data. This indicated that the overall library complexity was not notably impacted and, thus, also our V017 and V013 libraries remain suitable for read count analyses despite the significantly increased mean read pair duplication for the V013 miniaturization. Both miniaturization factor V017 and V013 have in common that all reagents were diluted to match the V025 volume. We are not aware of any published research examining similar aspects and can therefore only speculate about a potential link between dilution and read pair duplication. If there is a relationship, the most likely reason would be a decreased effectiveness of the cDNA synthesis step in the library preparation prior to the PCR amplification. The decreased effectiveness could have led to a slight decrease in library complexity which in return could have caused an increase in duplicates created by the PCR amplification.

We evaluated non-random RNA fragmentation as a potential cause for a library bias and tested for it by looking at the nucleotide base ratios on both sides of the library fragmentation site. Only at eight positions, one or more bases were significantly ($P < 0.001$) changed between un-miniaturized and miniaturized samples for both RIL DR8 and DR9 (Figure S4). All eight positions were found in the comparison between V013 and V100. We

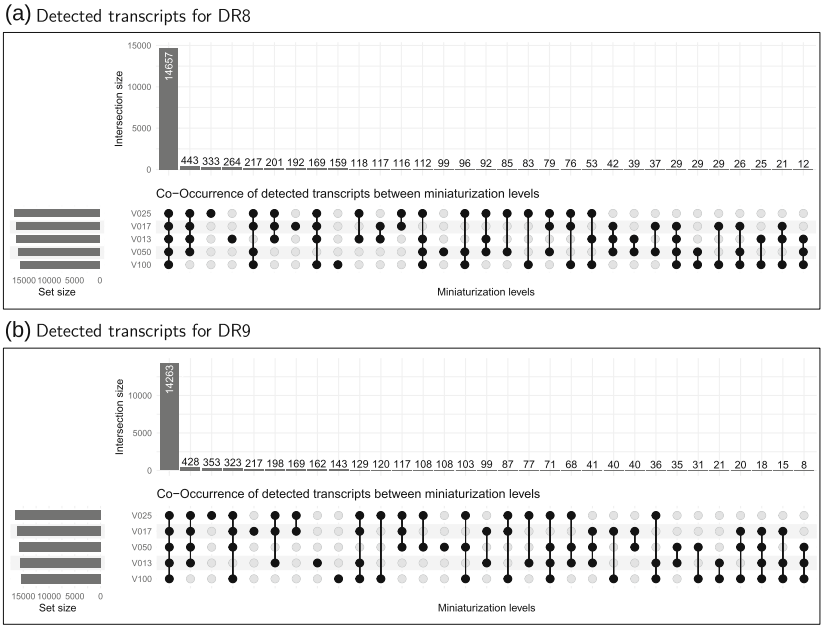


Figure 6 Co-occurrence of detected transcripts between miniaturization levels. Overlap of detected transcripts was calculated based on a subset of 2 million reads of the two recombinant inbred lines (RIL) DR8 (a) and DR9 (b), respectively.

suspect that the significant changes are caused by changes in sequence composition in samples with a high number of duplicates. However, these differences did not lead to a 5′ gene body coverage bias in any sample with no considerable difference between the different miniaturization levels (Figure 2d). All the above-mentioned results indicate that our modification to the library preparation workflow did not negatively impact the library complexity and did not introduce biases.

Quality control: Increased variability among replicates

Two aspects that were significantly ($P < 0.001$) changed by the miniaturization were (1) the proportion of raw read duplications and (2) the rate of multi-aligned reads. First, the increase in raw read duplications for the miniaturizations V025, V017 and V013 compared to V50 and V100 coincided with an increased variability among replicates within these miniaturization levels (Figure 3a). The number of raw read duplications can change based on differences in transcriptome composition between genotypes (Bansal, 2017). However, both RIL that were included in all examined miniaturization levels showed the same trend (Table S3) and, therefore the genotype was only partially able to explain the observations.

It is unclear if the remaining duplicated reads were introduced by sampling biological duplicates based on the redundancy of mRNA molecules or were technical duplicates created during PCR amplification. Expanding the workflow to include unique molecular identifiers (UMIs) would make the distinction between these two cases possible, but also increases the cost (Kivioja *et al.*, 2011). However, UMIs are also reported to increase the accuracy of transcript quantification, which could justify the added cost under certain circumstances (Fu *et al.*, 2018). Results of previous studies suggested that the number of usable reads can be increased by up to 40% by deduplication using UMIs (Collins *et al.*, 2015; Fu *et al.*, 2018; Girardot *et al.*, 2016). However, the selection of commercial kits that use adapters with UMI is limited when the number of multiplexed samples is bigger than 96. At the time of writing, we are not aware of any dual indexed UMI adapter kits that would have been compatible with our library preparation kit and have the capability of a 384-sample multiplex.

The raw read duplicates were further characterized by estimating the multiplication levels of the duplicated sequences. The number of unique duplicates, defined as the number of unique sequences that were duplicated, was consistent between all samples and miniaturization levels and did not reflect the variability in raw read duplicates (Figure S5a). This indicated that the sample-dependent copy number increase of a relatively small number of sequences was responsible for the observed variability.

Secondly, the number of properly paired unique alignments was reduced and the variability increased for the miniaturization levels V025 and above compared to V050 and V100 (Figure 3b). This observation cannot be fully explained by the difference between RILs (Figure 3d). In addition, the rate of multi-aligned reads and the rate of raw read duplications were strongly correlated on a sample-by-sample basis (Figure 3c). The connection between these two measures was not clear to us, which is why we further investigated their nature.

The origin of the increased variability

We began by thoroughly examining the relationship between the rate of multi-aligned reads and the rate of raw read duplications on a read-by-read basis, comparing the rate of raw read duplications between the subset of only multi-aligned reads and all reads in the total data set. The average proportion of duplicated reads was increased by 36% in the subset of multi-aligned reads and included up to 73% of all duplicates. This observation indicated that many multi-aligned reads are also duplicated reads (Figure S6a,b). To better understand the source of multi-aligned reads, we investigated their biological origin. One biological reason for the occurrence of multi-aligned reads are gene duplication events (Deschamps-Francoeur *et al.*, 2020). Therefore, we have examined genomic features that are classically connected to genome duplications and repetitiveness such as rRNA, histone gene family and transposable elements (Deschamps-Francoeur *et al.*, 2020; Magadum *et al.*, 2013; Rooney and Ward, 2005).

When checking the positional overlap of our reads with TEs in barley, the overlap in the subset of multi-aligned reads was higher than in the total data set (Figure S9a). This indicated that at least a part of the variance of the multi-alignment rate across samples can be explained by a sample-dependent increase in TEs. At the same time, a GO term enrichment analysis between the subset of multi-aligned reads and the total data set indicated a reduction in GO terms associated with retrotransposon activity. The same GO term enrichment analysis resulted in no significant ($P > 0.05$) changes in GO terms connected to the synthesis, modification or regulation of histones (Figure S7). Importantly, only a small proportion of the subset of multi-aligned reads was included in the analysis, because most of the reads could not be assigned to a gene (Figure S6a). This favours a non-gene-related explanation for the variance of multi-alignment rates like TE read or rRNA contamination.

From a technical perspective, the contamination with rRNA would be the most plausible explanation for the variance of multi-alignment rates. We tested for rRNA by using available rRNA annotation data in *Hordeum vulgare* to search for rRNA sequences in our multi-aligned read subset. The mean overall alignment rate for the read subset was considerably higher than for the total data set (Figure 4a). We could show that the increase in multi-alignments was strictly correlated with an increase in rRNA reads (Figure 4b). The rate of multi-aligned reads that could not be identified as rRNA was low (3.6%) and remained constant across all 94 samples (Figure 4c). Additionally, on average less than 2% of rRNA sequences were not aligning multiple times. These observations strongly suggested that the increase in multi-alignments as well as the increase in raw read duplications in the miniaturizations V025, V017 and V013 compared to V050 and V100 was caused by rRNA contamination.

The origin and consequences of rRNA contamination

We speculate that the rRNA contamination is caused by incomplete separation during the mRNA capture process at the beginning of the library preparation. This is the first and one of the most crucial magnetic bead separation steps, which are in our experience the most error-prone steps and very susceptible to the decrease in volume caused by the miniaturization (Figure S11).

Depending on the planned usage of the RNA sequencing data, an increased rRNA sequence content leads to a decrease in useful sequencing read output. This in return requires the number of sequenced reads to be increased, which inflates the costs. When accounting for the highest difference in multi-aligned read rates that was observed in our study for a single sample with 35%, the increased sequencing depth added 6 Euro to the overall costs. This was considerably lower than the overall cost reduction of the miniaturized library preparation by 21 Euro realized in our study (Table 1). Therefore, even accounting for the highest possible rRNA sequence content detected in this study, the cost-saving potential of miniaturized library preparation remains attractive.

To prevent rRNA contamination, the poly-A mRNA capture method in our workflow could be replaced by an rRNA depletion step, which was shown to have higher success in removing rRNA than poly-A capture methods (Kumar et al., 2017). This would presumably decrease the number of rRNA molecules and consequently the number of raw read duplications and multi alignments in our data set. However, the addition of an rRNA depletion step would greatly reduce the cost saving potential of the workflow and therefore only rarely be a reasonable alternative. Alternatively, for reaction volumes below 30 µL that include magnetic beads, a special low-elution 96-well magnet plate can be used to increase the magnetic bond and enable more reliable molecule selection. When this study was performed, a low-elution 96-well magnetic plate was not available to us and we were using a standard magnetic plate (96S Super Magnet, Alpaqua Engineering, USA). This most likely caused the mRNA capture and therefore the rRNA exclusion success to vary when handling smaller volumes in the higher miniaturizations.

Table 1 Summary of workflow costs. The relative costs comparison for the RNA isolation, RNA library preparation (RNA library prep.) and RNA sequencing between all miniaturization factors. The costs are relative to a standard workflow defined as: RNA extraction using RNeasy Plant Mini Kit (Qiagen, Germany) and library preparation using TruSeq RNA Library Prep Kit v2 (Illumina, USA). In comparison, the miniaturization workflow used: TRIzol-96 RNA isolation, VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina library preparation. Both workflows aimed to sequence 10 million reads using the DNBSEQ-G400 platform. The sequencing depth was increased to compensate for multi-aligned reads according to the maximum multi alignment rate for each miniaturization factor. The multi alignment rate of the standard workflow was set to 0%

	Miniaturization factors			
Workflow	V100	V050	V025	V013
RNA isolation	31.47%	15.74%	15.74%	15.74%
RNA library prep.	30.66%	15.33%	7.66%	3.83%
RNA sequencing	111.95%	115.32%	135.41%	147.07%
Total	44.39%	32.19%	29.70%	28.72%

Workflow modifications

The here described workflow is compatible with state-of-the-art liquid handling automatization solutions. The labour and time intensive magnetic bead separation steps would greatly benefit from the integration of an automated liquid handler designed for working with 96-well plates. The steps that would benefit are: the poly-A tail capture, size selection and all library fragment clean-ups. The other steps would require a liquid handling device that is able to accurately transfer small volumes (e.g. positive displacement instruments, acoustic droplet ejection instruments). This would potentially enable miniaturization factors beyond the ones examined in this study.

When using different types of plant material, the requirements to the workflow could change. We see a potential need to adjust the RNA extraction method depending on the material used. We showed that our miniaturization workflow produces high-quality data with multiple types of extraction methods. While the extraction methods can slightly alter the overall costs of the workflow, it does not affect the validity of the miniaturization as a cost saving measure. After successfully extraction of high-quality total RNA, our preliminary data do not indicate that the later steps of the workflow will be affected by the type of input material.

Data application: Sequence variation and read counts

We picked two common use cases to examine the suitability of the RNA sequencing data generated with our workflow. First, the potential of the data sets to identify sequence variants and second, the ability to describe differential expression using read

counts. The capabilities of the data set to call sequence variation and the potential influence the miniaturization had on it, were investigated by comparing the mean number of detected SNPs. The total number of SNP for DR8 and DR9, respectively did not change significantly ($P > 0.05$) between miniaturizations (Table S5). This finding was in agreement with the observation made in the PCA we conducted using filtered imputed SNP data (Figure 5). The results showed a high percentage of the variance between samples that matched the genetic structure of the samples, while the miniaturization was not shown to explain any variance and therefore could not be shown to have a systematic effect on the generated data (Figure S10). This makes the workflow capable of detecting sequence variances, which then could be used for example, in genome-wide association studies (Rodriguez *et al*., 2020).

When further examining the read counts, between 85%–92% of the consensus transcripts (expressed at least once per miniaturization) were found in all miniaturizations and the overlap between V100 and V025 was >97% (Figure 6). This illustrates the high consistency between miniaturizations which was further underpinned by Pearson correlation coefficients of at least 0.9613 for the RIL DR8 and DR9 among the miniaturization levels (Table S1). The correlation coefficient was even higher when comparing RNA extraction and library preparation replicates for both RIL (Table S2). These observations suggested that the miniaturization did not affect the ability to capture the transcriptome and all levels are capable of being used for comparative read count analyses for example, in differential gene expression (Cantalapiedra *et al*., 2017; Kintlová *et al*., 2021).

Conclusions

With minimal adjustments to an established commercial RNAseq library preparation protocol, we were able to manually miniaturize the library preparation by a factor of up to 1:8. This leads to cost savings of up to 54.5% compared to the same library preparation protocol without miniaturization and up to 86.1% compared to the gold standard. The rigorous quality control analysis of the resulting sequencing data and its application did not result in any indication of biases or inaccuracies caused by the library miniaturization. All libraries created in this study can be considered high quality and are ready to be used in a wide range of projects. The observed rRNA contamination did not affect the quality of the library itself and can be addressed by workflow adjustments, for example, using low elution volume magnetic plates. As shown by our cost projections even the potential efficiency decrease caused by the increase in multi-aligned reads did not change the fact that the method proposed here is an uncomplicated way to reduce the cost of RNAseq library preparation without a long set-up time or the need for scarcely available lab automation equipment. While in our study only a single commercial RNAseq library preparation kit was evaluated, we are confident that the general principle of miniaturization as well as observed unbiased results can be applied to a wide range of kits.

Materials and methods

Genetic material

Our study was based on five barley RIL from three HvDRR subpopulations (Casale *et al*., 2021). The HvDRR population was developed from pairwise crosses among 23 diverse parental inbreds (Weisweiler *et al*., 2019) using the double round robin (DRR) mating design (Stich, 2009).

Experimental design

The experimental design was set up such that it is possible to statistically test the effect of the miniaturization, RNA extraction method, number of PCR cycles and degree of plant tissue grinding on library complexity and biases (Figure 1). A total of 96 samples were examined. Five different genotypes were used: from HvDRR sub-population #28 line #33 (DR1), #46 (DR12) and #57 (DR10), from HvDRR sub-population #27 line #40 (DR8) and from HvDRR sub-population #13 line #29 (DR9). The libraries were prepared without miniaturization (1:1, V100), the miniaturization levels 1:2 using 50% of all reagents (V050), 1:4 using 25% of all reagents (V025), 1:6 using 17% of all reagents (V017) and 1:8 using 13% of all reagents (V013). Two of the five RIL (DR8 and DR9) were included in all miniaturization levels to allow orthogonal comparisons. Depending on the miniaturization level, two or three RNA extraction replicates and library preparation replicates each was present. RNA extraction replicates were distinct RNA extractions and library preparations using the same plant material. Library preparation replicates were distinct library preparations using the same extracted RNA. For the miniaturization levels 1:4, 1:6 and 1:8, the library of line DR8 was amplified using 8 PCR cycles and 10 PCR cycles. For the remaining miniaturization levels, only 10 PCR cycles were used. DR1 libraries were only prepared using 1:4 miniaturization but with plant material being either coarsely or finely ground as the basis for RNA isolation. Finally, 1:4 miniaturization libraries for the lines DR10 and DR12 were prepared using different RNA extraction methods.

Plant cultivation

A total of 15 seeds from each of the five RIL were sterilized with sodium hypochlorite (13%) for 10 min. All seeds of a single RIL were placed in a rectangular (12 × 12 cm) Petri dish between two sheets of filter paper (12 × 12 cm) supplied with tap water. The seeds were lined up in the lower half of the Petri dish on top of the water-soaked first filter paper. A second water-soaked filter paper was placed above the seeds, starting 1 cm above the seeds so that the lower third of the paper was reaching out of the Petri dish (Figure S12). This ensured that both filter paper sheets do not dry out over time. The Petri dishes were then stacked and placed in a vertical orientation in a plant tray (40 × 60 cm). This way, the space requirements per RIL were minimized and when filling the tray with water (approx. 3 L) the seedlings can grow for more than 7 days without further maintenance. The seedlings were cultivated for 7 days in a reach-in growth chamber under the following conditions: 70% relative humidity, 16 h of light (6:00–22:00), 22 degrees (day)/20 degrees (night) and light intensity of 400 μmol m^{-1} s^{-1}. The time of day for the cultivation start and the harvest was similar (within 2 h) for all samples. The RIL DR8, DR9 and DR10 were cultivated at the same time, while DR1 and DR12 were each cultivated at different dates.

Plant material processing

The whole seedlings of one Petri dish were harvested by transfer into a collection tube and immediately freezing them in liquid nitrogen. Afterward, they were ground by using steel beads and a paint shaker until the plant material was powdery (fine). Additionally, half of the plant material from DR1 was only

ground until it was flaky (coarse). Afterward, 50 mg or 100 mg of plant material was transferred into different collection tubes depending on the extraction method and stored at −80 degrees until RNA extraction started.

RNA extraction

Three different total RNA extraction methods were applied. First, a custom Phenol:Chloroform extraction optimized for high throughput extraction in 96-well plates (Box *et al*. (2011); Phenol:Chloroform, P:C). Second, a TRIzol reagent (Thermo Fisher Scientific) based RNA extraction following the manufactures instructions (TRIzol). For the third RNA extraction, TRIzol reagent (Thermo Fisher Scientific) was used in a 96-well format with an adapted protocol (TRIzol-96). The input plant material and all reagents for the TRIzol-96 extraction were halved compared to the standard protocol. The final washing step in 75% ethanol was repeated one time to assure that all the remaining phenol was removed. All other steps were executed as proposed by the manufacturer. The first two methods used 100 mg and the third method 50 mg of frozen fresh plant material as input. The total RNA concentration was quantified using a NanoPhotometer NP 80 (Implen, Germany). All samples, except the TRIzol-96 extractions, were evaluated using the Fragment Analyzer (Agilent).

Library preparation

The mRNA was selected based on a poly-A tail mRNA capture method (Vazyme, China) using 1 μg total RNA as input. The full-length mRNA library was constructed using the VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina (Vazyme, China). We miniaturized the kits by reducing the reagent volume to 50% (V050), 25% (V025), 17% (V017) and 13% (V013) of the original. The reaction volume for V017 and V013 were kept at 25% of the original volume to avoid pipetting volumes below 1 μL. The remaining reaction volume was filled up with RNase-free water, which resulted in a dilution for the miniaturizations V017 (1:1.5) and V013 (1:2). Size selection and clean-up were performed using magnetic DNA Clean Beads (Vazyme, China). In that step, the reaction volume and reagents were reduced to the respective miniaturization level with the exception of V017 and V013 for which the V025 reaction volume and reagents were used. Apart from these changes, the manufacturer's protocol was followed aiming for 250–450 bp long inserts. The 96 separate libraries were prepared in a 96-well plate with each miniaturization level occupying two to four columns. The order of the columns was randomized starting with the library preparation. The costs of our experimental workflow were compared to a gold standard which includes RNA extraction using RNeasy Plant Mini Kit (Qiagen, Germany) and library preparation using TruSeq RNA Library Prep Kit v2 (Illumina, USA).

Sequencing, read processing and alignment

The sequencing was performed by BGI on the DNBSEQ-G400 platform. All 96 samples were pooled and a total of 1.42 billion 150 bp paired-end reads were sequenced with an average of 14.8 million read pairs per sample. Two samples did have less than 2 million reads sequenced and were excluded from all further analyses. Both samples were from RIL DR8, RNA extraction replicate #1 and miniaturization levels V013 and V017, respectively. Various quality statistics of the raw sequencing reads were calculated using FastQC (Andrews, 2019) and afterward trimmed with trimmomatic (ILLUMINACLIP:TruSeq3-PE:2:30:10:1:TRUE SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36) (Bolger *et al*., 2014). The trimmed reads were then aligned to the Morex V3 reference sequence (Mascher, 2019) using Hisat2 (−no-softclip −max-seeds 1000) (Kim *et al*., 2019).

Alignment against rRNA reference libraries

In an effort to learn about the origin of multi-aligned reads, a read subset was created including only reads flagged as multi-mapped in the primary alignment against the reference sequence Morex. The total data set and the subset of multi-aligned reads were then aligned against two different rRNA reference libraries to estimate the percentage of reads originating from rRNA.

The rRNA reference libraries were created using HISAT2 without exon and splice site information. The sequences were searched for and downloaded as .fasta files from the RNA Central Expert Database using the following search criteria: (1) *Hordeum vulgare* subsp. *vulgare* rRNA (search term: taxonomy: "112509" AND rna_type: "rRNA") (1399 sequences) and (2) Ensembl Plant database rRNA (search term: rna_type: "rRNA" AND expert_db: "Ensembl Plants") (14 880 sequences). The much larger Ensembl Plant library includes rRNA sequences from many different plant species (e.g. *Arabidopsis thaliana*, *Oryza sativa* Japonica, *Triticum aestivum*, *Hordeum vulgare*) was used to evaluate the *Hordeum vulgare* library capability to create a comprehensive rRNA sequences alignment.

Variant calling and SNP analyses

Variant calling was performed using bcftools mileup (filter: -q 20 -Q 20) and call functions (Li *et al*., 2009). The variants were filtered based on the QUAL score (≥10), the median read depth per sample (≥5) and the total depth per variant (≥30). The raw variant call data was imputed using Beagle 5.4 (Browning *et al*., 2021) based on standard settings without a reference sequence. In the resulting SNP data set, all monomorphic and triallelic SNP and all SNP with more than 30% heterozygosity were removed. The remaining heterozygote SNP were set to NA and afterward median imputed. The SNP count comparisons of filtered and imputed variants used mean SNP counts of each miniaturization and were based on a 2 million read subset.

Read count analysis

The sorted and filtered alignments were then used to determine the read count per gene with the help of htseq count (−mode union) (Anders *et al*., 2015). The read counts were filtered and the Trimmed Mean of the M-values (TMM) method was used to apply a between-sample normalization using the R package edgeR (Robinson *et al*., 2009). The mean Pearson correlation was calculated for all pairwise library replicate combinations within each RNA extraction replicate. Each DR8 and DR9 library replicate was averaged for each miniaturization level. For calculating the correlation between the RNA extraction replicates, the mean of the read counts across the library preparation replicates was used. Afterward, the mean of the correlations was calculated for both DR8 and DR9. Pearson correlations were also calculated between miniaturizations using the mean read counts of all available replicates. Read counts were not only calculated for the total data set but also for the subset of multi-aligned reads. Here we had to allow for non-unique alignments to be included using the '–nonunique all' option of the htseq count function. The number of detected transcripts was calculated based on raw read counts

of 2 million read subsets of all samples. For the estimation of the number of consensus transcripts, transcripts with read counts below 10 were set to 0 in all samples. Afterward, the number of transcripts present was counted for all combinations of miniaturization levels.

Gene body coverage

In order to estimate the relative gene body coverage, a non-random subset of gene-associated reads was evaluated based on their relative position within the gene. We divided each transcript into 100 equally sized windows and counted the number of overlapping reads for each window. Each read was allowed to be counted multiple times. This analysis was performed for all expressed transcripts and afterward, the mean number of reads per sample for each window was calculated. The means were adjusted to accommodate for varying numbers of total reads per sample and rescaled to the range [0, 1] using general minimum and maximum read counts. Because of limitations of the Rsamtools R-library, only transcripts in the first 536 870 912 bases of each chromosome were included in the analysis.

Additional data analyses

To evaluate, if miniaturization leads to non-random fragmentation, the rate of each of the four nucleotide bases was calculated for the first nine bases before and after a fragmentation site. The first base of each forward read was defined as the first base after a fragmentation site. Unless the read start was equal to a transcript start. The first nine bases after the fragmentation site were therefore the first nine bases of a forward read and the nine bases before a fragmentation site were the last nine bases of a reverse read. A GO term enrichment analysis was conducted between the total data set and the subset of multi-aligned reads. Statistical differences were calculated using the Fisher exact test. The *P*-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

The rate of TE reads in the total data set and the subset of multi-aligned reads was estimated by calculating the rate of reads that overlap with genome positions annotated as TEs. Significant differences between groups (e.g. miniaturization levels) were assessed with *post hoc* Tukey's honestly significant difference tests. A PCA of the filtered, imputed variants across all samples was performed. A similar analysis was conducted on the TMM-normalized read counts. All metrics ascertained during the general sequencing data processing were aggregated using multiQC (Ewels *et al*., 2016). The significance threshold for all statistical tests in this study was set to 0.05.

Author contributions

The study was conceptualized and designed by C.A and B.S.; K.K. and T.W advised the laboratory experiments performed by C.A; The data was analysed and interpreted by C.A and B.S; The manuscript was written by C.A. and edited by B.S., C.A., K.K. and T.W; All authors read and approved the final manuscript.

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich-Heine-University Duesseldorf. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID:390686111). Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflict of interest.

References

Aigrain, L., Yong, G. and Quail, M.A. (2016) Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for illumina sequencing. *BMC Genom.* **17**, 458.

Alberti, A., Belser, C., Engelen, S., Bertrand, L., Orvain, C., Brinas, L., Cruaud, C. *et al*. (2014) Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genom.* **15**, 912.

Alpern, D., Gardeux, V., Russeil, J., Mangeat, B., Antonio, C., Meireles-Filho, A., Breysse, R. *et al*. (2019) BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* **20**, 12.

Anders, S., Pyl, P.T. and Huber, W. (2015) Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Andrews, S. (2019) *Fastqc*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G. *et al*. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J. *et al*. (2018) Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 12.

Bansal, V. (2017) A computational method for estimating the PCR duplication rate in dna and RNA-seq experiments. *BMC Bioinform.* **18**, 113–123.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Box, M.S., Coustham, V., Dean, C. and Mylne, J.S. (2011) Protocol: A simple phenol-based method for 96-well extraction of high quality RNA from arabidopsis. *Plant Methods* **7**, 7.

Browning, B.L., X.T., Zhou, Y. and Browning, S.R. (2021) Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890.

Cantalapiedra, C.P., García-Pereira, M.J., Gracia, M.P., Igartua, E., Casas, A.M. and Contreras-Moreira, B. (2017) Large differences in gene expression responses to drought and heat stress between elite barley cultivar scarlett and a spanish landrace. *Front. Plant Sci.* **8**, 647.

Casale, F., Van Inghelandt, D., Weisweiler, M., Li, J. and Stich, B. (2021) Genomic prediction of the recombination rate variation in barley – a route to highly recombinogenic genotypes. *Plant Biotechnol. J.* **20**, 676–690.

Chung, Y.S., Choi, S.C., Jun, T.H. and Kim, C. (2017) Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Hortic. Environ. Biotechnol.* **58**, 425–431.

Collins, J.E., Wali, N., Sealy, I.M., Morris, J.A., White, R.J., Leonard, S.R., Jackson, D.K. *et al*. (2015) High-throughput and quantitative genome-wide messenger RNA sequencing for molecular phenotyping. *BMC Genom.* **16**, 1–13.

Dabney, J. and Meyer, M. (2012) Length and gc-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern dna sequencing libraries. *Biotechniques*, **52**, 87–94.

Deschamps-Francoeur, G., Simoneau, J. and Scott, M.S. (2020) Handling multi-mapped reads in RNA-seq. *Computational and Structural. Biotechnol. J.* **18**, 1569–1576.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048.

Foley, J.W., Zhu, C., Jolivet, P., Zhu, S.X., Lu, P., Meaney, M.J. and West, R.B. (2019) Gene expression profiling of single cells from archival tissue with laser-capture microdissection and smart-3seq. *Genome Res.* **29**, 1816–1825.

Fu, Y., Wu, P.H., Beane, T., Zamore, P.D. and Weng, Z. (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genom.* **19**, 1–14.

Girardot, C., Scholtalbers, J., Sauer, S., Shu Yi, S. and Furlong, E.E.M. (2016) Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinform.* **17**, 1–6.

Harris, P.N.A. and Wailan Alexander, M. (2021) Beyond the core genome: Tracking plasmids in outbreaks of multidrug-resistant bacteria. *Clin. Infect. Dis.* **72**, 421–422.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D. *et al.* (2016) CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 77.

Hou, Z., Jiang, P., Swanson, S.A., Elwell, A.L., Nguyen, B.K.S., Bolin, J.M., Stewart, R. *et al.* (2015) A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* **5**, 9570.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S. (2012) Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828.

Jaeger, B.N., Yángüez, E., Gesuita, L., Denoth-Lippuner, A., Kruse, M., Karayannis, T. and Jessberger, S. (2020) Miniaturization of smart-seq2 for single-cell and single-nucleus RNA sequencing. *STAR Protocols* **1**, 100081.

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., Coulée, M. *et al.* (2021) RNA-seq data for reliable snp detection and genotype calling: Interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. *Front. Genet.* **12**, 1104.

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.

Kintlová, M., Vrána, J., Hobza, R., Blavet, N. and Hudzieczek, V. (2021) Transcriptome response to cadmium exposure in barley (hordeum vulgare l.). *Front. Plant Sci.* **12**, 1359.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74.

Kong, S.L., Li, H., Tai, J.A., Courtois, E.T., Poh, H.M., Lau, D.P., Haw, Y.X. *et al.* (2019) Concurrent single-cell RNA and targeted DNA sequencing on an automated platform for comeasurement of genomic and transcriptomic signatures. *Clin. Chem.* **65**, 272–281.

Kumar, R., Ichihashi, Y., Kimura, S., Chitwood, D.H., Headland, L.R., Peng, J., Maloof, J.N. *et al.* (2012) A high-throughput method for illumina RNA-seq library preparation. *Front. Plant Sci.* **3**, 202.

Li, H. (2021) Single-cell RNA sequencing in drosophila: Technologies and applications. *Wiley Interdiscip. Rev. Dev. Biol.* **10**, e396.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li, H., Kun, W., Ruan, C., Pan, J., Wang, Y. and Long, H. (2019) Cost-reduction strategies in massive genomics experiments. *Mar. Life sci. Technol.* **1**, 15–21.

Linderman, M.D., Nielsen, D.E. and Green, R.C. (2016) Personal genome sequencing in ostensibly healthy individuals and the peopleseq consortium. *J. Pers. Med.* **6**, 14.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. and Ravikesavan, R. (2013) Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161.

Mardis, E.R. (2011) A decade's perspective on dna sequencing technology. *Nature*, **470**, 198–203.

Mascher, M. (2019) *Pseudomolecules and annotation of the second version of the reference genome sequence assembly of barley cv. morex [morex v2]*. https://doi.ipk-gatersleben.de:443/DOI/83e8e186-dc4b-47f7-a820-28ad37cb176b/d1067eba-1d08-42e2-85ec-66bfd5112cd8/2

Mayday, M.Y., Khan, L.M., Chow, E.D., Zinter, M.S. and DeRisi, J.L. (2019) Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. *PloS One*, **14**, e0206194.

McCombie, W.R., McPherson, J.D. and Mardis, E.R. (2019) Next-generation sequencing technologies. *Cold Spring Harb. Perspect. Med.* **162**, e59.

McNulty, S.N., Mann, P.R., Robinson, J.A., Duncavage, E.J. and Pfeifer, J.D. (2020) Impact of reducing DNA input on next-generation sequencing library complexity and variant detection. *J. Mol. Diagn.* **22**, 720–727.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Varela, A.Á., Batlle, E. *et al.* (2020) Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755.

Mildrum, S., Hendricks, A., Stortchevoi, A., Kamelamela, N., Butty, V.L. and Levine, S.S. (2020) High-throughput minitaturized RNA-seq library preparation. *J. Biomol. Tech.* **31**, 151–156.

Mora-Castilla, S., Cuong To, Vaezeslami, S., Morey, R., Srinivasan, S., Dumdie, J.N., Cook-Andersen, H. *et al.* (2016) Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* **21**, 557–567.

Pallares, L.F., Picard, S. and Ayroles, J.F. (2019) Tm3′seq: A tagmentation-mediated 3′ sequencing approach for improving scalability of rnaseq experiments. *G3: Genes—Genomes—Genetics*, **10**, 143–150.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2016) The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 1–11.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G. and Sandberg, R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1100.

Piskol, R., Ramaswami, G. and Li, J.B. (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009) edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodriguez, M., Scintu, A., Posadinu, C.M., Yimin, X., Nguyen, C.V., Sun, H., Bitocchi, E. *et al.* (2020) Gwas based on RNA-seq snps and high-throughput phenotyping combined with climatic data highlights the reservoir of valuable genetic diversity in regional tomato landraces. *Gene*, **11**, 1–25.

Romero, I.G., Pai, A.A., Tung, J., Gilad, Y., Romero, I.G., Pai, A.A., Tung, J. *et al.* (2014) RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42.

Rooney, A.P. and Ward, T.J. (2005) Evolution of a large ribosomal RNA multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5084–5089.

Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J. *et al.* (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods*, **12**, 323–325.

Shomroni, O., Sitte, M., Schmidt, J., Parbin, S., Ludewig, F., Yigit, G., Zelarayan, L.C. *et al.* (2022) novel single-cell RNA-sequencing approach and its applicability connecting genotype to phenotype in ageing disease. *Sci. Rep.* **12**, 1–14.

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. and Mikkelsen, T.S. (2014) *Characterization of directed differentiation by high-throughput single-cell RNA-seq. bioRxiv*. 003236.

Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656.

Stich, B. (2009) Comparison of mating designs for establishing nested association mapping populations in maize and arabidopsis thaliana. *Genetics*, **183**, 1525–1534.

Tegally, H., San, J.E., Giandhari, J. and de Oliveira, T. (2020) Unlocking the efficiency of genomics laboratories with robotic liquid-handling. *BMC Genom.* **21**, 12.

Vahrenkamp, J.M., Szczotka, K., Dodson, M.K., Jarboe, E.A., Soisson, A.P. and Gertz, J. (2019) Ffpecap-seq: A method for sequencing capped rnas in formalin-fixed paraffin-embedded samples. *Genome Res.* **29**, 1826–1835.

Wang, T., Liu, Y., Ruan, J., Dong, X., Wang, Y. and Peng, J. (2021) A pipeline for RNA-seq based eQTL analysis with automated quality control procedures. *BMC Bioinform.* **22**, 1–18.

Weisweiler, M., De Montaigu, A., Ries, D., Pfeifer, M. and Stich, B. (2019) Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mrna sequencing and their power to predict phenotypic traits. *BMC Genom.* **20**, 10.

Wetterstrand, K.A. (2021) *The cost of sequencing a human genome*. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Fragment Analyzer results for total RNA.
Figure S2. Fragment Analyzer results for the library pool.
Figure S3. Sequencing read quality control.
Figure S4. Calculation of nucleotide base distribution around fragmentation sites.
Figure S5. Overview of different multiplication levels.
Figure S6. Duplication rates and rRNA alignment.
Figure S7. GO term enrichment between the subset of multi-aligned reads and the total data set.
Figure S8. Investigation of the no feature (NF) read origin in the subset of multi aligned reads.
Figure S9. Investigation of the transposable element (TE) read origin in the subset of multi-aligned reads.
Figure S10. Principal component analysis based on RNA sequencing data of the five recombinant inbred lines (RIL).
Figure S11. Evaluation of the impact of miniaturization on magnetic bead selection steps
Figure S12. Illustration of the cultivation system.
Table S1. Correlation of read counts between miniaturizations.
Table S2. Correlation of read counts between library preparation and RNA extraction replicates.
Table S3. Mean and standard deviation (SD) of the raw read duplication rate.
Table S4. Mean and standard deviation (SD) of the alignment statistics.
Table S5. Comparison of SNP counts between miniaturizations.

5 Assessment of genomic prediction capabilities of transcriptome data in a barley multi-parent RIL population

This manuscript was accepted for publication in Theoretical and Applied Genetics in July 2025.

Authors:
Christopher Arlt, Delphine van Inghelandt, Jinquan Li, and Benjamin Stich.

Own contribution: First author. I created the data, performed the data analysis, and wrote the manuscript.

Assessment of genomic prediction capabilities of transcriptome data in a barley multi-parent RIL population

Christopher Arlt[1,2], Delphine van Inghelandt[2,3], Jinquan Li[4] and Benjamin Stich[2,4,5,6,*]

[1] Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany.

[2] Institute for Breeding Research on Agricultural Crops, Julius Kühn Institute (JKI) - Federal Research Centre for Cultivated Plants, 18190 Sanitz, Germany

[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany.

[4] Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

[5] Cluster of Excellence on Plant Sciences (CEPLAS), 40225 Duesseldorf, Germany.

[6] University of Rostock, 18051 Rostock, Germany.

*Corresponding author: Tel: +49 38209 45201; Fax: +49 38209 45222; E-Mail: Benjamin.Stich@julius-kuehn.de; ORCID: https://orcid.org/0000-0001-6791-8068

Additional author information

- Christopher Arlt; E-Mail: christopher.arlt@julius-kuehn.de;
  ORCID: https://orcid.org/0000-0001-8735-5549
- Delphine van Inghelandt; E-Mail: delphine.inghelandt@julius-kuehn.de;
  ORCID: https://orcid.org/0000-0002-2819-843X
- Jinquan Li; E-Mail: j.li@strube.net;
  ORCID: https://orcid.org/0009-0004-8171-1281

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich-Heine-University Duesseldorf.

Running head

Transcriptome-based prediction in barley

Keywords

Genomic selection, Plant breeding, Quantitative genetics, Barley, Transcriptomics, RNA sequencing.

Key message

Low-cost and high-throughput RNA sequencing data for barley RILs achieved GP performance comparable to or better than traditional SNP array datasets when combined with parental whole genome sequencing SNP data.

ABSTRACT

The field of genomic selection (GS) is advancing rapidly on many fronts including the utilization of multi-omics datasets with the goal to increase prediction ability (PA) and to become an integral part of an increasing number of breeding programs ensuring future food security. In this study, we used RNA sequencing (RNA-Seq) data to perform genomic prediction (GP) on three related barley RIL populations investigating the potential of increasing PA by combining genomic and transcriptomic datasets, adding whole genome sequencing (WGS) SNP data, functional parameter filtering, and empirical quality filtering. Our RNA-Seq data were generated cost-efficiently using small footprint plant cultivation, high-throughput RNA extraction, and library preparation miniaturization. We also examined the depth of the sequencing as an additional cost-saving measure. We used five-fold cross-validation to evaluate the PA of the gene expression dataset, the RNA-Seq SNP dataset, and the consensus SNP dataset between the RNA-Seq and parental WGS data, resulting in PAs between 0.73 and 0.78. The consensus SNP dataset performed best, with five out of eight traits performing significantly better compared to a 50K SNP array, which served as a benchmark. The advantage of the consensus SNP dataset was most prominent in the inter-population predictions, in which the training- and validation-set originated from different RIL sub-populations. We could therefore not only show that RNA-Seq data alone are able to predict various complex traits in barley using RIL, but also that the performance can be further increased by WGS data for which the public availability will steadily increase.

INTRODUCTION

The continuous increase in the global population and the per capita crop demand create the need for agricultural expansion and advancements in plant breeding (Lenaerts et al., 2019; Tilman et al., 2011). The ecological impact of land clearing due to agricultural intensification is already substantial and will increase further (Zabel et al., 2019; Burney et al., 2010). This makes improving crop yield one of the most important tasks (Burgess et al., 2023).

The improvement of agriculturally important quantitative traits by plant and animal breeding was achieved for thousands of years without knowledge of genetics principles but simply based on artificial human selection on phenotypes (Purugganan and Fuller, 2009; Pourkheirandish and Komatsuda, 2007; Wright, 2005). This changed in recent decades with the increase in knowledge and capabilities in genetics, which in turn was possible by dramatic progress on genotyping and sequencing approaches. This enabled the development and application of marker-assisted selection (MAS) (Dekkers and Hospital, 2002; Lande and Thompson, 1990; Fernando and Grossman, 1989). The challenge is that most agriculturally relevant traits are quantitative traits controlled by many genes with each having only a small effect (Glazier et al., 2002; Mackay, 2001), while MAS is most effective when large-effect loci contribute to the trait of interest (Heffner et al., 2009). With the increasing availability of high density genome-wide genetic marker data, genomic selection (GS) was introduced as a method to estimate breeding values using all available marker data (Meuwissen et al., 2001) and not only those that were significantly associated with the trait. GS alleviated the downside of classical MAS when trying to predict complex quantita-

tive traits with many small-effect loci (Zhao et al., 2014) and increased the rate of genetic gain. While GS was first studied in the context of animal breeding, it was later adopted by plant breeders (Heffner et al., 2009; Zhong et al., 2009; Bernardo and Yu, 2007). The reduction in costs to produce marker data using next-generation sequencing techniques such as genotyping by sequencing (GBS) further increased the popularity of GS (Bhat et al., 2016).

Over the last decade, advances in GS methods have led to increased prediction abilities. For example, utilizing high-throughput phenotyping data as predictors increased the performance of multivariate GS models (Rutkoski et al., 2016). Additionally, traditional GS models were expanded to multi-trait GS models (Tsai et al., 2020; Lyra et al., 2017; Jia and Jannink, 2012). Furthermore, multi-environmental GS models increased the predictability in multiple studies (Hu et al., 2023; Li et al., 2019). The most recent advances are the inclusion of deep learning, machine learning, and artificial intelligence in the GS workflow (Sandhu et al., 2021; Montesinos-Lopez et al., 2021; Bayer et al., 2021; Washburn et al., 2020; Harfouche et al., 2019). This is an ongoing field of research and is not yet fully explored, with studies showing limitations or at least the need for adjustments to current implementations (Ubbens et al., 2021).

While the core of most GS models is a relationship matrix derived from genomic marker data, in the field of multi-omic GS additional sources of information like transcriptome and metabolome data can be used as predictors. The transcriptome is a promising predictor, bridging the gap between the genome and the trait (Azodi et al., 2020). Quantified gene expression can be captured using a microarray or mRNA sequencing (RNA-Seq). RNA-Seq data are more versatile than microarrays

as from such data not only gene expression information can be extracted but also sequence variants in the portion of the genome covered by the sequenced RNA (Azodi et al., 2020; Weisweiler et al., 2019). Because both aspects are rarely considered together and to our knowledge no research is available that evaluates functional or quality based sub-setting of sequence variants from RNA-Seq data, the full potential of RNA-Seq datasets for GS is not yet sufficiently studied.

Within the field of plant breeding, several studies have previously shown that including transcriptome and metabolome data (multi-comics) in GS models has the potential to increase prediction capabilities. For example, Michel et al. (2021) added incomplete RNA-Seq data to complete genomic marker data to assess disease resistance phenotypes in wheat. In addition, a multi-omics prediction study in oat used transcriptomic and metabolomic data to compare single-environmental trials and multi-environmental trials (Hu et al., 2021). Guo et al. (2016) was able to successfully combine transcriptomic and metabolomic data with genomic markers from diverse maize inbred lines to increase predictability in GS. Data using multi-omics GS in barley, on the other hand, are extremely limited and only recently started to emerge (Wu et al., 2022).

Almost all previous studies focused on diversity panels to validate the capabilities of their GS model (Hu et al., 2021; Michel et al., 2021; Westhues et al., 2019; Schrag et al., 2018; Westhues et al., 2017; Guo et al., 2016). However, in plant breeding programs, half-sib or full-sib families are typically used from which the most appropriate progenies are selected. Nevertheless, to the best of our knowledge, no earlier study evaluated the potential of multi-omic GS models in this context.

In this study, we explore the capabilities of low-cost RNA-Seq data to perform genomic prediction (GP) on three connected spring barley RIL populations with 237 individual lines for eight agriculturally important traits each measured in up to seven environments. Based on this dataset, the potential to increase the performance of the GP model by combining genomic and transcriptomic data, functional parameter filtering, and a novel empirical two-step quality filtering approach is evaluated. Lastly, we examine multiple optimization parameters that could lead to cost and time savings without sacrificing prediction ability.

MATERIALS AND METHODS

Genetic material

The HvDRR population was developed from pairwise crosses among 23 diverse parental inbreds (Weisweiler et al., 2019) using the double round-robin (DRR) mating design (Stich, 2009). Our study was based on 237 recombinant inbred lines (RIL) from three HvDRR sub-populations (Casale et al., 2022) that were derived from pairwise crosses among parental inbreds Spratt-Archer, HOR8160, and Unumli-Arpa (Fig. 1). The three sub-populations were designated in the following as HvDRR13 (65 RIL), HvDRR27 (92 RIL), and HvDRR28 (80 RIL).

Plant cultivation for RNA extraction

All RIL were cultivated in a randomized augmented incomplete block design. A block consisted of 24 samples, including 21 RILs and all three parents as controls (Fig. 1). The cultivation workflow was identical to that described previously (Arlt et al., 2023). In summary, for each RIL, 15 seedlings were cultivated in vertically stacked square Petri dishes for seven days in a reach-in growth chamber under the following conditions: 70% relative humidity, 16 hours of light (6:00 - 22:00), 22 degrees (day) / 20 degrees (night), and light intensity about 400 $\mu mol\ m^{-2}\ s^{-1}$ (Fig. S1). The time of day for planting and harvesting were similar (within two hours) for all samples. All plants in the same block were processed simultaneously.

RNA extraction

The seedlings were harvested as a whole, immediately frozen, and ground. From 50 mg of plant material, total RNA was extracted using TRIzol reagent (Thermo Fisher, USA). The manufacturer's protocol was adapted as described below to fit a

96-well format and to use less reagent (Arlt et al., 2023). The input plant material and all the reagents for extraction were halved. The final washing step in 75% ethanol was repeated one additional time to ensure that all the remaining phenol was removed. All other steps were performed as proposed by the manufacturer. The total RNA concentration was quantified using a NanoPhotometer NP 80 from Implen (Germany). A total of 33 extractions were randomly selected for evaluation using the Fragment Analyzer (Agilent, USA).

Library preparation

The mRNA was selected based on a poly-A tail mRNA capture method (Vazyme, China) using $1\mu g$ total RNA input. The full-length mRNA library was constructed using the VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina kit from Vazyme (China). We miniaturized the procedure by reducing the volume of reagent to 25% of the original amount (Arlt et al., 2023). Size selection and cleanup was performed using magnetic DNA Clean Beads (Vazyme, China). Apart from the miniaturization, the manufacturer's protocol was followed aiming for 250-450bp long inserts.

Sequencing and read processing

Sequencing was performed by BGI on the DNBSEQ-G400 platform. All 284 samples were pooled and a total of 3.70 billion 150 bp paired end reads were sequenced with an average of 13.0 million read pairs per sample (Arlt and Stich, 2022). Various quality statistics of raw sequencing reads were evaluated using fastQC (Andrews, 2019) and afterwards trimmed with trimmomatic (ILLUMINACLIP:TruSeq3-PE:2:30:10:1:TRUE SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36) (Bolger et al., 2014). The trimmed reads were then aligned to the Morex V3 reference

sequence (Mascher, 2019) using Hisat2 (–no-softclip –max-seeds 1000) (Kim et al., 2019).

Sequence variant calling

Variant calling was performed using the following three datasets: RNA sequencing, SNP array, and whole genome sequencing (WGS) data. (1) For all RNA-Seq datasets, variant calling was performed using the bcftools mpileup (filter: -q 20 -Q 20) and call function (Li et al., 2009). All duplicated genotypes were united using the major allele as consensus. All variants with missing parental genotype information were excluded. Afterwards, the data was cleaned by setting all heterozygous alleles to NA, due to the near complete homozygosity of all inbred lines used in this study. Additionally, all RIL alleles inconsistent with parental alleles were set to NA. The missing data was median imputed for each RIL population independently. The resulting dataset was designated in the following as SNP^{Total}_{RNAseq}. To test the impact of quality filtering on a less strictly cleaned dataset SNP^{Raw}_{RNAseq} was created, which did not set heterozygous / inconsistent allele calls to NA and did not remove variants with missing parental data. (2) Already existing SNP array data for the same RIL were included in this study (SNP^{Total}_{Array}). The data were generated using the Illumina 50K iSelect SNP array for barley (Bayer et al., 2017). The SNP array dataset was filtered as described by Casale et al. (2022). (3) Additionally, all available RNA-Seq SNP were intersected with WGS SNP data from the parental inbreds (Weisweiler et al., 2022), selecting only variants that were present in both datasets and therefore building a consensus subset (SNP^{Total}_{WGS}). For SNP^{Total}_{WGS}, missing SNP data were imputed using parental WGS SNP data with Beagle (Browning et al., 2021). For all datasets described above, genetic differentiation G_{st} was calculated according to Nei (1973).

RNA-Seq SNP function filtering

The function of the SNP in the SNP_{RNAseq}^{Total} dataset was determined using the SnpSift and SnpEff programs (Cingolani et al., 2012a,b). We used the SNP functional information to filter the total dataset and create two new subsets. The first SNP functional subset contained only SNP within the 5'UTR and 3'UTR within a 5 kb distance to the coding region (SnpSift.jar filter "(ANN[*].EFFECT has 'upstream_gene_variant') || (ANN[*].EFFECT has 'downstream_gene_variant')") and was designated in the following as $SNP_{RNAseq}^{Reg.}$. The second SNP function subset excluded all synonymous SNP, only selecting missense variant SNP in the coding region (SnpSift.jar filter "ANN[*].EFFECT has 'missense_variant'") and was designated in the following as SNP_{RNAseq}^{CDS}. Missing data was median imputed and all monomorphic markers were excluded.

RNA-Seq SNP quality filtering

The following quality filtering parameters were used to filter SNP_{RNAseq}^{Total}: read depth (DP), minor allele frequency (MAF), quality score (QUAL), and number of samples without data (NS). The filtering was based on the information from the vcf file. After determining the minimum and maximum values of SNP_{RNAseq}^{Total} for the four criteria, the full spectrum was divided into 21 segments based on relative filtering strength (from 0% to 100% in 5% steps). For each step, the results of the genomic prediction cross-validation were analyzed to determine the relative filter strength that performs best for each criterion.

In a last step, the subsets with the best filtering performances were combined creating the inclusive marker intersects, which were evaluated for their prediction ability. The results were compared to a standard SNP filtering procedure (missing

rate < 20%, MAF > 0.05) (Atanda et al., 2022; Wen et al., 2018) and the prediction abilities were tested for their significance using pairwise t-tests. The best performing combination was selected and the data were cleaned and imputed analogously to the procedure of SNP^{Total}_{RNAseq}. The resulting quality-filtered RNA-Seq SNP subset was designated in the following as SNP^{QC}_{RNAseq}.

Read count calculation

The expression per transcript was determined for all samples with the help of ht-seq count (–mode union) (Anders et al., 2015). The Trimmed Mean of the M-values (TMM) method was used to apply a between-sample normalization using the R package edgeR (Robinson and Oshlack, 2010). All transcripts in which at least 2% of the samples had non-zero read counts were included in the total expression dataset. The original read counts (GE^{Total}_{RNAseq}) were filtered using a counts per million (cpm) threshold which maximizes the number of differentially expressed genes (DEG) resulting in the filtered expression dataset ($GE^{Filter.}_{RNAseq}$). The method used is similar to the DEseq2 filtering approach (Love et al., 2014). DEG was calculated for each of the available genotypes (all RILs of the three populations and the parental inbreds) and compared to the general mean. The dispersion was estimated by edgeR based on the replicated parental inbreds. An additional dataset was created that included only the DEG (GE^{DEG}_{RNAseq}). For all expression subsets, a per transcript normalization was applied to correct for any block effects on the expression levels. We used the following linear mixed model for each transcript: $y_{ijk} = \mu + G_i + B_j + \varepsilon_{ijk}$, with the genotype effect (G_i, fixed), block effect (B_j, random), and an error term (ε_{ijk}). Afterwards, the adjusted entry means were calculated using the emmeans package (Searle et al., 1980). This normalization method was designated in the following as EMM normal-

ization. The GP performance was evaluated using, only EMM normalization, only TMM normalization, both normalization methods together (EMM first, TMM first), and with un-normalized read counts.

Phenotypic datasets

In this study eight phenotypic traits were considered (Tab. 1). All phenotypic information was collected in field experiments using an augmented row-column design in which the RIL were planted with a single replicate. The parental inbreds were used as checks with multiple replicates. The awn and spike length data were collected between 2019 and 2021 in five different environments. Flowering time and plant height were part of field experiments from 2017 to 2019 in seven environments (Cosenza et al., 2024). Grain length, grain width, grain area, and thousand grain weight were measured between 2017 and 2019 in four different environments (Shrestha et al., 2022). The adjusted entry means for the phenotypic values were calculated using the following mixed linear model: $y_{ijk} = \mu + G_i + E_j + (G \times E)_{ij} + \varepsilon_{ijk}$, with the genotype effect (G_i, fixed), environment effect (E_j, random), and the genotype-environment interaction effect ($(G \times E)_{ij}$, random), as well as an error term (ε_{ijk}). The heritability for all phenotypic traits was calculated as $H^2 = \sigma^2_G/(\sigma^2_G + \bar{\nu}/2)$, where $\bar{\nu}$ was the mean variance of difference between two adjusted entry means (Piepho and Möhring, 2007).

Genomic prediction

We used all eleven datasets described above (Tab. 2) to predict the adjusted entry means for all eight traits. The values of all expression and sequence variant datasets were converted into z-scores ($\mu = 0$, $\sigma = 1$) for error normalization. Afterwards, additive relationship matrices were calculated for each of the datasets: $G = \frac{W^* {W^*}^T}{m}$,

where W^* was the z-score matrix of the feature measurements W, W^{*^T} was the transposed z-score matrix and m the number of features per dataset. We used the following genomic best linear unbiased prediction (GBLUP) model (VanRaden, 2008): $y = \mu + Zu + \varepsilon$, where y was the vector of adjusted entry means of the examined trait, μ the general mean, Z the incidence matrix of genotypic effects, and u the vector of genotypic effects that are assumed to be normally distributed with $u \sim N(0, G\sigma_u^2)$, in which G denotes the relationship matrix between inbreds and σ_u^2 the genetic variance. In addition, ε is the vector of residuals following a normal distribution $\varepsilon \sim N(0, I\sigma_e^2)$. This approach was used across all three segregating populations. Multiple predictor datasets were included in the model by calculating the weighted average between the relationship matrices G (Wu et al., 2022). The weight of each matrix was equal, if not otherwise stated. The training and testing population was simulated using 5-fold cross-validation and 50 repetitions.

Genomic prediction: Intra and inter-population analyzes

In addition to the prediction across the three RIL populations, various intra- and inter-population analyses were performed. For the intra- and inter-population genomic prediction, fixed size randomized samples were used as training and testing set with 200 repetitions. The relationship matrices and GP model were identical to the above described 5-fold cross-validation method. We used multiple subsets from a single RIL population with a fixed number of individuals as training and testing set to evaluate the intra-population prediction ability of our data.

We created a training and testing set from the same RIL population. The number of individuals for each set was the same for all RIL populations. The inter-population prediction ability was tested by using one of the three RIL populations as training set

and a second as testing set. The number of individuals included in both sets was kept constant. Intra- and inter-population results were compared to the GP performance using cross-population training and testing sets.

Genomic prediction: Reduced sequencing depth

To evaluate the impact of reduced sequencing depth on GP performance, we tested 13 sequencing depth subsets. The number of reads per sequencing depth subset ranged from 10K to 7M. The reads to include were randomly selected from all uniquely mapped reads of the alignment data. Samples which did have less uniquely mapped reads than required were excluded. To ensure comparability, the same samples were included across all datasets for which GP was performed. All other steps of the data processing workflow were not changed. We tested the impact on RNA-Seq sequencing variant datasets for three different filtering methods creating the reduced sequencing depth equivalents of SNP^{Total}_{RNAseq}, $SNP^{Stn.}_{RNAseq}$, and SNP^{QC}_{RNAseq} for each of the 13 sequencing depths. Similarly, the expression datasets were evaluated including all transcripts (GE^{Total}_{RNAseq}), filtered transcripts ($GE^{Filter.}_{RNAseq}$), and DEG (GE^{DEG}_{RNAseq}).

RESULTS

Characterization of gene expression data and their GP performance

RNA-Seq data was used to characterize the expression pattern of 237 RIL from the HvDRR13, HvDRR27, and HvDRR28 population and their parental inbreds Spratt-Archer, HOR8160, and Unumli-Arpa (Fig. 1). The number of detected transcripts per sample ranged from 22.1K to 30.9K (mean 28K) and did not show significant differences ($p > 0.05$) between the genetic material groups (Fig. 2A). The different segregating populations were clearly separated in the expression-based PCA (Fig. 2 B). The RIL within each population were arranged according to their phenotypic values (Fig. S2).

The total gene expression dataset (GE^{Total}_{RNAseq}) which included 42.6K transcripts was divided into two subsets: (1) transcripts with an average cpm > 0.39 ($GE^{Filter.}_{RNAseq}$; 37.7K transcripts) and (2) a subset only including DEG (GE^{DEG}_{RNAseq}; 7.3K transcripts)(Fig. 2C). We determined the cpm threshold for the $GE^{Filter.}_{RNAseq}$ dataset based on the maximum number of significant DEG detected (Fig. S3). This resulted in a cpm threshold lower than the standard edgeR threshold of about 1.7. The standard filtering approach of edgeR resulted in $<$ 6K DEG.

We measured gene expression per transcript as well as per gene and detected a slight but significant increase ($p = 0.009$) in PA for the former in GE^{Total}_{RNAseq} (Fig. S4A). We also compared different normalization methods. For the un-normalized cpm and the TMM normalized cpm, the GP performance did not differ significantly ($p > 0.05$), while the EMM normalized GP results were significantly lower ($p = 0.003$) compared to the PA of the un-normalized results (Fig. S4B-C).

The PA of the GE^{Total}_{RNAseq} set was with an average 0.68 across all traits, the lowest of all gene expression subsets. The GP performance of $GE^{Filter.}_{RNAseq}$ was slightly higher (0.69) but both were outperformed by GE^{DEG}_{RNAseq} which had an average prediction ability of 0.73 (Fig. 2D). The PA differed among the traits. Flowering time was in all subsets the most difficult to predict resulting in an average prediction ability between 0.59 and 0.67. Ear length was the trait most accurately predicted with an average prediction ability between 0.78 and 0.82.

Prediction ability of sequence variant datasets

In addition to the expression data, sequence variant calling was performed for the RNA-Seq dataset comprising 148K variants after cleaning and imputation (SNP^{Total}_{RNAseq}). The SNP data was further subset based on variant function annotation, creating a set that only included variants in regulatory regions of genes ($SNP^{Reg.}_{RNAseq}$) and a set that only included missense variants in gene coding sequences (SNP^{CDS}_{RNAseq}). The latter led to an increase in PA of 0.013 compared to SNP^{Total}_{RNAseq} (Fig. 3). A comparable increase in PA was achieved by filtering the RNA-Seq variants using quality criteria.

In this study, two quality-filtered subsets were evaluated. First, our own empirical quality filtering subset designated in the following as SNP^{QC}_{RNAseq} and second, the subset $SNP^{Stn.}_{RNAseq}$ which is based on previously published filtering criteria (Atanda et al., 2022; Wen et al., 2018). We used the following quality criteria for the empirical quality filtering workflow: read depth (DP), minor allele frequency (MAF), missing data (NS), and quality score (QUAL). More than 20 quality filtering subsets of SNP^{Total}_{RNAseq}, ranging from no filtering to maximum filtering strength, were created and their GP performance was evaluated (Fig. S5). For each of the quality

filtering criteria, the best performing SNP^{Total}_{RNAseq} subset was selected and combined with one or more of the remaining subsets (Fig. 4A). Most combinations showed a slight increase in PA and the best performing combination (designated in the following as SNP^{QC}_{RNAseq}) was 0.002 higher than the subset of best performing single criteria of the combination (not significant: $p > 0.05$). SNP^{QC}_{RNAseq} included 42.5K variants, utilized MAF and QUAL as quality filter criteria, and performed with a 0.015 PA increase considerably better than SNP^{Total}_{RNAseq}. A slightly lower uplift was achieved by $SNP^{Stn.}_{RNAseq}$. However, when comparing the GP results of $SNP^{Stn.}_{RNAseq}$ and SNP^{QC}_{RNAseq} for each of the traits, SNP^{QC}_{RNAseq} performed significantly better ($p \leq 0.025$) for half of them (Fig. 4B). A PCA was performed using the SNP^{QC}_{RNAseq} subset and a separation between the three parental inbreds and RIL populations was observed as expected (Fig. S6A). Each of the eight traits showed clear differences among the parental inbreds that caused phenotypic variance in the corresponding segregating populations (Fig. S6B-I).

Lastly, WGS data from the parental inbreds were used to create an optimized WGS imputed RNA-Seq consensus variant dataset (SNP^{Total}_{WGS}) achieving a PA of 0.776. This constitutes a 0.033 increase compared to SNP^{QC}_{RNAseq} and 0.009 compared to the SNP array dataset. SNP^{Total}_{WGS} showed the highest PA averaged across all traits. Based on SNP^{Total}_{WGS}, two function-based subsets ($SNP^{Reg.}_{WGS}$ and SNP^{CDS}_{WGS}) were created, but none of them had a positive impact on the GP results.

GP performance of combined datasets

We selected the best performing datasets from these four main groups: gene expression (GE^{DEG}_{RNAseq}), RNA-Seq SNP (SNP^{QC}_{RNAseq}), RNA-Seq/WGS consensus SNP (SNP^{Total}_{WGS}), and array SNP (SNP^{Total}_{Array}). We tested the combined GP performance

by merging two or more of these datasets (Fig. 5). For that, the additive relationship matrices calculated from each of the individual datasets were averaged, and the resulting matrix was used for GP. The combination of two datasets resulted in an average increase in PA of 0.01. Three of the six combined matrices were able to outperform the best dataset included in the combination (Fig. 5A). The average performance of all triple dataset combinations (0.77) was increased by 0.02 compared to the single datasets and by 0.01 compared to the double dataset combinations. Combining all four datasets resulted in an average PA of 0.78, which is 0.02 higher compared to the average single dataset GP and slightly better than the triple dataset combination average. However, none of the combinations showed a significantly higher PA ($p > 0.05$) than the best individual dataset included in the combination.

Examining each trait separately revealed a significant ($p < 0.01$) increase in GP performance for at least one of the traits in multiple combined datasets (Fig. 5B). The best performing combination consisted of SNP^{QC}_{RNAseq} and SNP^{Total}_{Array}, as it significantly increased PA for three traits ($p \leq 0.016$) and only decreased significantly ($p < 0.001$) once. Combining SNP^{Total}_{WGS} and GE^{DEG}_{RNAseq} with equal weight resulted in a significant decrease in prediction ability for all eight traits. We also tested for significant differences compared to SNP^{Total}_{Array} (Tab. S1). SNP^{Total}_{WGS} significantly outperform SNP^{Total}_{Array} in five of eight traits. Combining SNP^{Total}_{WGS}, SNP^{QC}_{RNAseq} and GE^{DEG}_{RNAseq} significantly increased half of the traits ($p \leq 0.025$), which was the best-performing combination without including SNP^{Total}_{Array}. We further evaluated the combination including GE^{DEG}_{RNAseq} and SNP^{Total}_{WGS} by testing a range of weighted combinations (Fig. 5C). We started with only the GE^{DEG}_{RNAseq} dataset and gradually increased the percentage of weight contribution SNP^{Total}_{WGS} until only the SNP^{Total}_{WGS} data remained.

The results of the trait flowering time showed a significant increase (p = 0.007) in GP performance for an unbalanced combination with a 90% weight contribution by SNP^{Total}_{WGS}. For all other traits, the combinations resulted in lower PA.

We quantified the differences of the same four datasets by calculating Pearson's correlation coefficient of the additive relationship matrices (Fig. S7). The correlation between SNP^{QC}_{RNAseq} and SNP^{Total}_{WGS} was the highest with a correlation coefficient of 0.98. The correlation of SNP^{Total}_{Array} with SNP^{Total}_{WGS} ($r = 0.91$) and SNP^{QC}_{RNAseq} ($r = 0.89$) showed slightly lower level of similarity. The gene expression matrix (GE^{DEG}_{RNAseq}) was more dissimilar compared to the remaining datasets ($r = 0.59 - 0.61$). We split the additive relationship matrices into segments based on the populations to calculate the correlation for all populations and inter-population combinations separately. We focused on the correlation between SNP^{Total}_{Array} and the remaining three datasets (Fig. S8) and showed for all three matrix combinations that the correlation coefficient was lower for the intra-population covariances than for the inter-population covariances. Additionally, large differences in similarity were shown for intra-population covariances.

Comparison of intra- vs. inter-population GP

To compare the GP performance using population-specific combinations of training (TS) and validation sets (VS), we had to adjust the validation procedure to account for the variable population sizes. Therefore, we switched from a 5-fold cross-validation scheme to a random subset validation with fixed training and testing set sizes. We established a baseline by performing cross-population GP using samples of multiple populations in TS and VS (Fig. 6A). The prediction ability with a low TS size (20) varied between 0.47 (GE^{DEG}_{RNAseq}) and 0.58 (SNP^{Total}_{Array}). These results could be

improved by 0.25 for GE^{DEG}_{RNAseq} and by 0.18 for SNP^{Total}_{Array} when increasing TS size to 170.

We then separated the populations and tested intra-population GP performance (TS = VS population) and inter-population GP performance (TS ≠ VS population) with a fixed TS size of 50 (Fig. 6B). For most datasets and comparisons, the intra-population GP results were slightly lower than cross-population results, with only SNP^{QC}_{RNAseq} and SNP^{Total}_{WGS} showing slightly higher performance (0.02) for HvDRR28. The prediction abilities of the inter-population GP results were lower compared to the intra-population GP results. GP performance depended on the combination of populations tested and resulted in prediction abilities ranging from 0 to 0.29. The worst performing combination between HvDRR28 (TS) and HvDRR13 (VS) was best predicted by GE^{DEG}_{RNAseq}.

Although the overall ranking between the datasets did not change between intra- and inter-population GP (Fig. S9), the relative differences between them varied (Fig. 6C). SNP^{Total}_{WGS} performed 5% better than SNP^{QC}_{RNAseq} in intra-population GP and the difference increased to 25% in inter-population GP. The difference in PA between SNP^{Total}_{WGS} and SNP^{Total}_{Array} increased from $< 1\%$ in intra-population GP to almost 3% better performance of SNP^{Total}_{WGS} in inter-population GP. The difference in performance between SNP^{QC}_{RNAseq} and GE^{DEG}_{RNAseq} was largest in intra-population GP (13%) but decreased to 3% in inter-population GP aligning the performance of both datasets.

Effect of sequencing depth reduction on GP performance

We tested the GP performance of 13 sub-sampled datasets with reduced sequenc-

ing depths from 7M to 10K reads using RNA-Seq SNP and gene expression data including 155 samples with sufficient reads (Fig. S10) and from 3M to 10K reads including 229 samples (Fig. 7). The GP performance of the three RNA-Seq SNP datasets was significantly ($p < 0.001$) affected by a reduction in sequencing depth (Fig. 7A). For SNP^{Total}_{RNAseq}, the prediction ability decreased only insignificantly ($p > 0.05$) for all subsets that started with more than 400K uniquely mapped reads (200K for $SNP^{Stn.}_{RNAseq}$ and SNP^{QC}_{RNAseq}). Reduction in sequencing depth below 100K reads impacted the GP performance considerably more. SNP^{Total}_{RNAseq} showed higher prediction ability values than $SNP^{Stn.}_{RNAseq}$ below 50K variants, while the quality filtering dataset SNP^{QC}_{RNAseq} increased the performance in all subsets. For the three gene expression datasets, performance started to decline earlier, with only the 2M and 3M subsets showing no significant differences between each other ($p > 0.05$) (Fig. 7B). GE^{DEG}_{RNAseq} performed worst in the 10K subset and only exceeded the prediction ability values of the other two datasets at sequencing depths over 100K.

DISCUSSION

Filtering of RNA-Seq datasets to unravel the full GP potential

We started with independently evaluating the GP performance of sequence variants and gene expression datasets derived from RNA-Seq data in a cross-population scenario. For most applications, the accuracy of RNA-Seq gene expression data is increased by the inclusion of replicates to counter act the technical errors introduced by experimental factors (Schurch et al., 2016). However, we expect that RNA-Seq will only find its way into routine plant breeding programs, if it can be applied without replications and as such it was examined in our study. Therefore, we normalized the gene expression values per transcript based on a linear model including a block effect to adjust the read counts (EMM) (Fig. S4B). While in 88% of all transcripts in GE^{DEG}_{RNAseq} the block effect was significant ($p < 0.05$), adjusted gene expression values negatively affected GP performance and led to a significant PA decrease ($p < 0.001$). Further research is necessary to learn whether a block normalization can be helpful in normalizing gene expression data in a controlled environmental setting similar to this study.

In addition to the EMM normalization, we tried to account for differences in transcriptome composition or sequencing lane effects by applying the per-sample normalization TMM implemented in edgeR (Robinson and Oshlack, 2010; Dillies et al., 2013), but it did not significantly ($p > 0.05$) change the prediction ability of GE^{DEG}_{RNAseq} (Fig. S4C). We assume that this is because all examined genotypes were partially related, limiting differences in transcriptome composition. Furthermore, the library pool was evenly distributed across all lanes.

Although normalization did not lead to an uplift in GP performance, we were able to increase PA by filtering for DEG. The resulting dataset GE^{DEG}_{RNAseq} was the best performing gene expression dataset across all traits with an average prediction ability of 0.73 (Fig. 2). Filtering also increased GP performance in the sequence variant datasets extracted from the RNA-Seq data (Fig. 3). In the end, the prediction based on sequence variation exceeded the prediction based on gene expression when averaged across all traits. This finding is in agreement with previous results (Azodi et al., 2020). This indicates a limited prediction potential of gene expression data which could have multiple reasons. Similarly to Azodi et al. (2020), we were using only seedling material for GP. Therefore, most spatiotemporal specific DEG were not captured in our data (Klepikova et al., 2016), which potentially limits GP performance. Conversely, creating trait-specific transcriptome data by focusing on the most relevant time point, developmental stage, and tissue could increase the ability to predict specific traits.This, however, is unlikely to be realistic in the context of commercial plant breeding programs. The overall number of transcripts detected was limited by the sequencing depth (Conesa et al., 2016). However, the sequencing depth in our study was chosen so that the related costs are comparable to that of genotyping using a SNP array. Ultimately, the prediction ability potential of transcriptome datasets can only be fully exploited by grouping and weighting individual genes, using prior knowledge of the trait and associated gene networks. Similar methods exist but have never been applied in this context (Zarringhalam et al., 2018). We are convinced that future research can further increase the GP performance of transcriptome data, while the costs of creating such datasets will remain comparable to the costs of an SNP array.

We evaluated the impact of function filtering of the unfiltered RNA-Seq based variant data (SNP^{Total}_{RNAseq}) based on two criteria: SNP in the 3'UTR and 5'UTR and nonsynonymous SNP in the CDS. In our study, both function filtering criteria achieved comparable GP performance while relying only on a fraction of the total variants, which also was previously observed (Li et al., 2022; Cappetta et al., 2021; Tan et al., 2017). While the 3'UTR and 5'UTR are known to be regulatory regions in which SNP can be associated with gene expression variation (Dossa et al., 2021; Zhang et al., 2019), in our study filtering missense SNP in the CDS resulted in a greater uplift of GP performance.

Quality filtering using the four quality criteria read depth, minor allele frequency, missing rate, and quality score were able to increase prediction ability compared to SNP^{Total}_{RNAseq}. The combination of multiple quality filter criteria further increased the GP results and led to the best-performing RNA-Seq sequence variant dataset SNP^{QC}_{RNAseq} (Fig. 4). Although our empirical quality filtering workflow did not result in a much larger average prediction ability compared to the commonly used filter settings (Atanda et al., 2022; Wen et al., 2018), for half of the traits evaluated in this study, PA was significantly increased ($p \leq 0.025$) by empirical quality filtering. We leveraged the information of the homozygosity of the RILs, the population structure, and the parental information to strictly clean and correct the sequence variant data, which reduced the impact of quality filtering on the GP performance. However, we observed that without strict cleaning using the before mentioned prior knowledge, the impact of quality filtering was much greater (Fig. S11). We applied the empirical quality filtering workflow to SNP^{Raw}_{RNAseq}, which resulted in a prediction ability of 0.75 which is comparable to that of SNP^{QC}_{RNAseq}. This observation suggests that our

empirical quality filtering workflow is therefore also able to provide good prediction abilities in situations with limited prior knowledge and, thus, might be relevant for other studies.

The last approach to improve the prediction ability that we examined was the exploitation of SNPs from parental inbred lines. In our study, the consensus dataset from the RNA-Seq and WGS variants (SNP_{WGS}^{Total}) was the overall best performing dataset for GP. SNP_{WGS}^{Total} outperformed SNP_{Array}^{Total} for five out of eight traits significantly ($p < 0.05$). Adding WGS data to improve GP performance was previously only examined in combination with SNP array data (Weber et al., 2024; Brøndum et al., 2015). Weber et al. (2024) showed insignificant improvements to the prediction accuracy and concluded that the WGS data were unable to add important information due to the relatedness of the sample groups and linkage disequilibrium between many WGS markers. In contrast, in our study the advantage of including WGS data is apparent and illustrates that for some predictor types, for instance, RNA-Seq and potentially also GBS datasets, the benefit from WGS data is more significant than for the already highly curated markers included in a SNP Array.

Cost optimization of GP using RNA-Seq

In addition to the filtering steps discussed above to improve the prediction ability, the selection of the optimal TS size is also important. In our study, increasing the TS size from 20 to 80 led to a drastically increased prediction ability (Fig. 6A). Increasing the TS to 170 only slightly increased the prediction ability further. The diminishing returns of TS size depends on the population structure and has been studied extensively before and therefore are not discussed further (Zhu et al., 2022; Bustos-Korts et al., 2016; Isidro et al., 2015; Rincent et al., 2012; Lorenz et al., 2012).

One further aspect to optimize the balance between prediction ability and cost is the sequencing depth. Our resampling simulations illustrated that the sequencing depth of RNA-Seq can be significantly reduced without sacrificing GP performance (Fig. 7). For SNP^{QC}_{RNAseq}, the reduction in GP performance was only marginal until the number of reads fell below 200K. This finding indicated that many sequence variants do not contain additional information, as they are part of the same haplotype block, and therefore reducing the number of reads and therewith sequence variants does not necessarily lead to a decreased trait prediction.

However, when reducing the sequencing depth further, we observed for the datasets that underwent a strong quality filtering a considerably reduction in PA. The $SNP^{Stn.}_{RNAseq}$ filtering resulted in < 100 sequence variants in the 10K read depth subset compared to the 6.9K marker in SNP^{Total}_{RNAseq}. Therefore, GP performance decreased as a consequence due to an insufficient number of polymorphic variants, which in turn led to an imprecise estimation of genetic relatedness. $SNP^{Stn.}_{RNAseq}$ and SNP^{Total}_{RNAseq} were outperformed by SNP^{QC}_{RNAseq} with 1.5K sequence variants at the same sequencing depth. These results illustrate the balance between the number of variants needed to achieve the necessary LD between variants and trait-coding polymorphisms and the selection of informative variants.

In comparison to the above outlined influence of a reduced sequencing depth on the prediction ability of sequence variant datasets, we observed a reduction of the prediction ability already at higher number of reads (around 3M) for the gene expression datasets. Even more pronounced was that the reduced number of reads led to a reduction in statistical power to detect DEG. For example, 400K reads were able to detect 76% of the transcripts of the total dataset, but only 22% of the DEG.

This ultimately resulted in less than 100 DEG in sequencing depth subsets below 20K, which led to dramatically reduced prediction abilities.

With reduced sequencing depth, the sequencing is most likely no longer the most expensive step of the project, and the remaining cost saving potential can be realized by adjusting the library preparation workflow. In this project, we used a thoroughly evaluated miniaturization workflow that reduced the costs of the library preparation by a factor of four (Arlt et al., 2023). At reduced sequencing depth, as discussed above, such approaches to reduce library preparation costs are even more crucial compared to our study.

Combining multiple datasets maximized GP performance

Besides the above described independent evaluations of the sequence variants and gene expression datasets, we also examined various scenarios on combining them. In detail, the overall best performing dataset SNP_{WGS}^{Total}, the gene expression dataset GE_{RNAseq}^{DEG}, the RNA-Seq sequence variance dataset SNP_{RNAseq}^{QC}, and the 50k SNP array dataset SNP_{Array}^{Total} were selected to be examined for their joint prediction ability. We observed a clear positive trend in PA when combining these datasets (Fig. 5). This was expected because their different information content complement each other.

However, not all combinations of individual datasets outperformed the best single predictor. The GP performance of all combinations was significantly decreased (< 0.05) for at least one trait. This highly trait-specific GP performance was in accordance with previous reports (Zhu et al., 2022; Michel et al., 2021) and can be explained by their differences in heritability and the number of small- and large-effect loci contributing to the trait. The high similarities between the sequence variant

datasets SNP^{QC}_{RNAseq}, SNP^{Total}_{WGS}, and SNP^{Total}_{Array} (Fig. S7) could explain the limited potential of their combinations to increase GP performance. The correlation coefficient indicated more differences between GE^{DEG}_{RNAseq} and SNP^{Total}_{Array} or SNP^{QC}_{RNAseq}, but combining them resulted only for two traits in significantly increased PA ($p \leq 0.014$). No clear increase in GP performance was observed by adding gene expression as an additional information layer to sequence variants, which was consistent with previous studies (Azodi et al., 2020; Guo et al., 2016).

In the first chapter of the discussion, we listed potential reasons for the limited PA potential of gene expression data and how to improve them. It is not clear whether such an improved transcriptome dateset would better complement or replace sequence variant data for the purpose of genomic prediction analyses. We also see a limited potential for datasets like ours to be further improved. For example, using logarithmic scaling of read counts increased the correlation coefficient between GE^{DEG}_{RNAseq} and the remaining datasets by 2-3% and slightly increased GP performance. This indicates that GP similar to differential expression analysis or co-expression analysis profits from the increased homoscedasticity of log-scaled expression values (Johnson and Krishnan, 2022; Love et al., 2014). Log-scaling could be especially beneficial when gene expression data are later combined with SNP data. Further research is needed to test this hypothesis.

Genetic diversity vs. relatedness: cross-, inter-, and intra-population prediction

Our multi-parent recombinant inbred line population allowed us to test multiple GP population designs and evaluate their performance using different input datasets (Fig. 6). Combining multiple populations in the TS (cross-population) increased the GP

performance in our study. This finding was in accordance with the results of Berro et al. (2019), but contrary to those of Lorenz et al. (2012) and illustrates that the ratio between population structure and diversity dictates whether the creation of a cross-population TS is beneficial.

The relative difference in PA between SNP_{WGS}^{Total} and the remaining datasets was higher for the inter-population GP vs. the intra-population GP (Fig. 6C). This could be explained by the increased marker count of SNP_{WGS}^{Total}, which could have led to a higher linkage between marker and QTL, which was shown to be more relevant when the genetic distance between TS and VS is increased (De Roos et al., 2009). Also noticeable, GE_{RNAseq}^{DEG} and SNP_{RNAseq}^{QC} performed similarly in inter-population GP, while in intra-population GP SNP_{RNAseq}^{QC} was clearly superior. This shift was mainly caused by a single inter-population combination that could be predicted by GE_{RNAseq}^{DEG} but not by SNP_{RNAseq}^{QC}.

SUMMARY

In this study, we evaluated various approaches to optimize the prediction potential of RNA-Seq datasets. Our results indicated that the RNA-Seq sequence data in combination with parental sequence data out-performed the SNP array data in five out of eight traits, showing a higher prediction ability overall. The outstanding performance of the consensus dataset combining WGS genomic variant information of the parental lines with the RNA-Seq sequence variant data is highly relevant. More and more diverse and for breeding relevant barley genotypes are sequenced, the amount of publicly accessible data will increase, allowing for comparable data combinations in the future to increase the GP potential in barley. The same is also true for other crop species.

When relying solely on RNA-Seq data, we showed that focusing on differentially expressed genes noticeably increased the prediction ability of the gene expression dataset but did not reach the prediction ability level of the sequence variant datasets. The GP performance of the RNA-Seq sequence variant data was substantially increased by strict data cleaning and quality filtering, but it only exceeded the prediction performance of the SNP array for one trait. Combining predictor datasets does not significantly ($p > 0.05$) increase PA averaged across all traits, but was able to significancy increase individual traits in half of the combinations tested. Therefore, the expected benefit of combing the RNA-Seq sequence variant and gene expression data did not materialize for our genetic material.

In this study, we showed that low-cost and high-throughput RNA-Seq data can achieve comparable or better GP performance than traditional SNP array datasets.

As the RNA-Seq data are more flexible, not relying on a pre-selected set of genomic variance, and allow for further decreased costs by adjusting the read depth and library preparation protocols, we see high application potential for RNA-Seq in plant breeding programs.

REFERENCES

S. Anders, P. T. Pyl, and W. Huber. Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu638. URL <GotoISI>://WOS:000347832300003https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287950/pdf/btu638.pdf. Ay8wu Times Cited:12890 Cited References Count:14.

Simon Andrews. Fastqc, 2019. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

C. Arlt, T. Wachtmeister, K. Köhrer, and B. Stich. Affordable, accurate and unbiased rna sequencing by manual library miniaturization: A case study in barley. *Plant Biotechnology Journal*, 21(11):2241–2253, 2023. ISSN 1467-7644. doi: 10.1111/pbi.14126. URL <GotoISI>://WOS:001050468100001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10579711/pdf/PBI-21-2241.pdf. U5ak0 Times Cited:1 Cited References Count:64.

Christopher Arlt and Benjamin Stich. Raw read rna-seq data (drrpop_13_27_28), 2022.

S. A. Atanda, V. Govindan, R. Singh, K. R. Robbins, J. Crossa, and A. R. Bentley. Sparse testing using genomic prediction improves selection for breeding targets in elite spring wheat. *Theoretical and Applied Genetics*, 135(6):1939–1950, 2022. ISSN 0040-5752. doi: 10.1007/s00122-022-04085-0. URL <GotoISI>://WOS:000776336900001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9205816/pdf/122_2022_Article_4085.pdf. 2e9im Times Cited:8 Cited References Count:53.

C. B. Azodi, J. Pardo, R. VanBuren, G. de los Campos, and S. H. Shiu. Transcriptome-based prediction of complex traits in maize. *Plant Cell*, 32(1):139–151, 2020. ISSN 1040-4651. doi: 10.1105/tpc.19.00332. URL <GotoISI>://WOS:000508875100014https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6961623/pdf/TPC_201900332R2.pdf. Ke9mw Times Cited:49 Cited References Count:54.

M. M. Bayer, P. Rapazote-Flores, M. Ganal, P. E. Hedley, M. Macaulay, J. Plieske, L. Ramsay, J. Russell, P. D. Shaw, W. Thomas, and R. Waugh. Development and evaluation of a barley 50k iselect snp array. *Frontiers in Plant Science*, 8, 2017. ISSN 1664-462x. doi: ARTN179210.3389/fpls.2017.01792. URL <GotoISI>://WOS:000413092400002https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2017.01792/pdf. Fj9jk Times Cited:172 Cited References Count:35.

P. E. Bayer, J. Petereit, M. F. Danilevicz, R. Anderson, J. Batley, and D. Edwards. The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome*, 14(3), 2021. URL <GotoISI>://WOS:000754723000004. Yy3xc Times Cited:0 Cited References Count:71.

R. Bernardo and J. M. Yu. Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082–1090, 2007. ISSN 0011-183x. doi: 10.2135/cropsci2006.11.0690. URL <GotoISI>://WOS:000247424200020. 181gc Times Cited:550 Cited References Count:27.

I. Berro, B. Lado, R. S. Nalin, M. Quincke, and L. Gutiérrez. Training population optimization for genomic selection. *Plant Genome*, 12(3), 2019. doi: ARTN1900

2810.3835/plantgenome2019.04.0028. URL <GotoISI>://WOS:000498828100014. Jq3da Times Cited:36 Cited References Count:69.

J. A. Bhat, S. Ali, R. K. Salgotra, Z. A. Mir, S. Dutta, V. Jadon, A. Tyagi, M. Mushtaq, N. Jain, P. K. Singh, G. P. Singh, and K. V. Prabhu. Genomic selection in the of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, 7, 2016. doi: ARTN22110.3389/fgene.2016.00221. URL <GotoISI>://WOS:000402686200001https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2016.00221/pdf. Ew7js Times Cited:176 Cited References Count:78.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170. URL <GotoISI>://WOS:000340049100004https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/pdf/btu170.pdf. Am7lq Times Cited:35103 Cited References Count:10.

B. L. Browning, X. W. Tian, Y. Zhou, and S. R. Browning. Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics*, 108(10):1880–1890, 2021. ISSN 0002-9297. doi: 10.1016/j.ajhg.2021.08.005. URL <GotoISI>://WOS:000705304300007https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8551421/pdf/main.pdf. We0ew Times Cited:160 Cited References Count:42.

R. F. Brøndum, G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*,

98(6):4107–4116, 2015. ISSN 0022-0302. doi: 10.3168/jds.2014-9005. URL https://dx.doi.org/10.3168/jds.2014-9005.

A. J. Burgess, C. Masclaux-Daubresse, G. Strittmatter, A. P. M. Weber, S. H. Taylor, J. Harbinson, X. Y. Yin, S. Long, M. J. Paul, P. Westhoff, F. Loreto, A. Ceriotti, V. L. R. Saltenis, M. Pribil, P. Nacry, L. B. Scharff, P. E. Jensen, B. Muller, J. P. Cohan, J. Foulkes, P. Rogowsky, P. Debaeke, C. Meyer, H. Nelissen, D. Inzé, R. K. Lankhorst, M. A. J. Parry, E. H. Murchie, and A. Baekelandt. Improving crop yield potential: Underlying biological processes and future prospects. *Food and Energy Security*, 12(1), 2023. ISSN 2048-3694. doi: 10.1002/fes3.435. URL <GotoISI>://WOS:000916112200001. 8p4fr Times Cited:11 Cited References Count:336.

J. A. Burney, S. J. Davis, and D. B. Lobell. Greenhouse gas mitigation by agricultural intensification. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):12052–12057, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0914216107. URL <GotoISI>://WOS:000279332300071https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2900707/pdf/pnas.200914216.pdf. 618dt Times Cited:700 Cited References Count:54.

D. Bustos-Korts, M. Malosetti, S. Chapman, B. Biddulph, and F. van Eeuwijk. Improvement of predictive ability by uniform coverage of the target genetic space. *G3-Genes Genomes Genetics*, 6(11):3733–3747, 2016. ISSN 2160-1836. doi: 10.1534/g3.116.035410. URL <GotoISI>://WOS:000388678000030https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5100872/pdf/3733.pdf. Ed2ly Times Cited:24 Cited References Count:73.

Elisa Cappetta, Giuseppe Andolfo, Anna Guadagno, Antonio Di Matteo, Amalia Barone, Luigi Frusciante, and Maria Raffaella Ercolano. Tomato genomic prediction for good performance under high-temperature and identification of loci involved in thermotolerance response. *Horticulture Research*, 8(1), 2021. ISSN 2662-6810. doi: 10.1038/s41438-021-00647-3. URL https://dx.doi.org/10.1038/s41438-021-00647-3.

F. Casale, D. Van Inghelandt, M. Weisweiler, J. Q. Li, and B. Stich. Genomic prediction of the recombination rate variation in barley - a route to highly recombinogenic genotypes. *Plant Biotechnology Journal*, 20(4):676–690, 2022. ISSN 1467-7644. doi: 10.1111/pbi.13746. URL <GotoISI>://WOS:000729149100001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8989500/pdf/PBI-20-676.pdf. 0i4kk Times Cited:11 Cited References Count:86.

P. Cingolani, V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, and X. Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Front Genet*, 3:35, 2012a. ISSN 1664-8021 (Electronic) 1664-8021 (Linking). doi: 10.3389/fgene.2012.00035. URL https://www.ncbi.nlm.nih.gov/pubmed/22435069https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3304048/pdf/fgene-03-00035.pdf. Cingolani, Pablo Patel, Viral M Coon, Melissa Nguyen, Tung Land, Susan J Ruden, Douglas M Lu, Xiangyi eng P30 ES006639/ES/NIEHS NIH HHS/ R01 DK071073/DK/NIDDK NIH HHS/ R01 ES012933/ES/NIEHS NIH HHS/ R21 ES021983/ES/NIEHS NIH HHS/ Switzerland 2012/03/22 Front Genet. 2012 Mar 15;3:35. doi: 10.3389/fgene.2012.00035. eCollection 2012.

P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu,

and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92, 2012b. ISSN 1933-6942 (Electronic) 1933-6934 (Print) 1933-6934 (Linking). doi: 10.4161/fly.19695. URL https://www.ncbi.nlm.nih.gov/pubmed/22728672https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/pdf/fly-6-80.pdf. Cingolani, Pablo Platts, Adrian Wang, Le Lily Coon, Melissa Nguyen, Tung Wang, Luan Land, Susan J Lu, Xiangyi Ruden, Douglas M eng P30 ES06639/ES/NIEHS NIH HHS/ R01 ES012933/ES/NIEHS NIH HHS/ P30 ES006639/ES/NIEHS NIH HHS/ R01 DK071073/DK/NIDDK NIH HHS/ R21 ES021983/ES/NIEHS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2012/06/26 Fly (Austin). 2012 Apr-Jun;6(2):80-92. doi: 10.4161/fly.19695.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17(1), 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8. URL https://dx.doi.org/10.1186/s13059-016-0881-8.

F. Cosenza, A. Shrestha, D. Van Inghelandt, F. A. Casale, P. Y. Wu, M. Weisweiler, J. Li, F. Wespel, and B. Stich. Genetic mapping reveals new loci and alleles for flowering time and plant height using the hvdrr population of barley. *J Exp Bot*, 2024. ISSN 1460-2431 (Electronic) 0022-0957 (Linking). doi: 10.1093/jxb/erae010. URL https://www.ncbi.nlm.nih.gov/pubmed/38330219https://watermark.silverchair.com/erae010.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL

39

_9Cf3qfKAc485ysgAAA5QwggOQBgkqhkiG9w0BBwagggOBMIIDfQIBADCCA3YGCSqGSI
b3DQEHATAeBglghkgBZQMEAS4wEQQMYhyInoq6vfwgCHwUAgEQgIIDR03o17-apzr1WQ
WOLNfPdfIMAv1LpewJaSKXOItLN4s095W68RpRR32_qK2FWjJdJMMp3UF4UrlBikX9GF
GHHt9bdjd0lxfXOTMc3tkcyITKwGD8rtmNejVlZaRv1ylJ6ZIfVm01qUufHtsj68YBUc
aeo4umX2neoR65A4V6TPeBWY94hjbeRjTZTxsJU_hQK27QYtIp42tqxzFraRJNuHbXCN
w9BRMGqUzrQPEuwXheh15FyCOKtqxF7aiOKWKqqa13skRisbJU2xCI943fMXD3CInsrZ
6YP2TDsphgJnBoPpwa9ZfwOy7Ot6sQq93Y7VkxxxQbrepvTbHJ4h_jwUlRw6ZL5obtyK
GVK0ktsSJPMMGoVJgvSSDnqf6gcUpBxTXfU_OOUE1rtBkotZMYrV-0-ZMm8DkwiMZi
sYNIYNqFuIlex64zWeq6KYREpHsLNpY8x7kPtNun2fKc0-Y8iBZ75S6eVuHbw_ftHbU
AOmDsxxv9BjLcsTvK5n6BDhv1VNTvdPElP--8M_odJkChOEuxMF9moW5ZRGxgW08ZbZ
7ieczaO3gFytL0hKt7RSsa5vm0J1tcUo0UGSQrXjCm0B-2lSuemd_bg2-SmbJsGHBU
atO5Q0dcb1vQawwH5vQqxZpn4TfhFPagIXeQ7EG8JCKNsfKry4uCwnaAclCk9qhueB
IFYsGmSEvqNW3QskgcHi2b8wmqSitnLfyh95z_Cw-xKmOfdMW4BKEq6hM36gNMjNM4
XJELX8VEwHZdpcM9UWkTAxk8CvTQ51ealBsmSKyIQrpsREnFCy1n33g6nJmepE6KnY
REz093c_Yt9DEAnyMBlfEe_P3YkhCJK4sQkwQTm-v0UOlQ0X6l33vBx05RNvIeQOevI
BKcMQiwDx8EsI3Qx55JDRY26DikJoDedA3dl6JUcV6-WFaIH1taxBLoLoTWkDF8tGF
Hcd4cwjGqOhM1HfWjw-eHolnN4fj0u0D3wKnGF-R-wIWvrFLFzau4NIHp7myqhjrlV
khvD_1bvmjeZtVKY1g8fzwG3IMf8fRd8_qesS7w0nOGfM4lib85QXA9JNsysTPIOXy
dBBWHt83HdjiVopEjk_QaaHDe_Z79tHF-DbLq. Cosenza, Francesco Shrestha, Asis Van Inghelandt, Delphine Casale, Federico A Wu, Po-Ya Weisweiler, Marius Li, Jinquan Wespel, Franziska Stich, Benjamin eng England 2024/02/08 J Exp Bot. 2024 Feb 8:erae010. doi: 10.1093/jxb/erae010.

A. P. W. De Roos, B. J. Hayes, and M. E. Goddard. Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553, 2009. ISSN 1943-2631.

doi: 10.1534/genetics.109.104935. URL https://dx.doi.org/10.1534/genetics.109.104935.

J. C. Dekkers and F. Hospital. The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet*, 3(1):22–32, 2002. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi: 10.1038/nrg701. URL https://www.ncbi.nlm.nih.gov/pubmed/11823788https://www.nature.com/articles/nrg701.pdf. Dekkers, Jack C M Hospital, Frederic eng Review England 2002/02/02 Nat Rev Genet. 2002 Jan;3(1):22-32. doi: 10.1038/nrg701.

M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013. ISSN 1467-5463. doi: 10.1093/bib/bbs046. URL <GotoISI>://WOS:000327436000001https://watermark.silverchair.com/bbs046.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAA5AwggOMBgkqhkiG9w0BBwagggN9MIIDeQIBADCCA3IGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMlZqwTHiW8UWubXPLAgEQgIIDQ2B-tv5V5IKmZmNo35DGZ1OExpoeX6XAn_Jw3PFTmPwhGsnEIKFs_KNCb7SmMyE8SAhGiZsYPfOZ6bE_KNhSkts1wV4c1OSaU6CtKeLZWBWG37_KVFqyJcpYGb0pN3zJgL6fv8Nba8sGlUAQkoefJ8gCCp2UkCIqX-FSz_M79gTrG9fQKBtCrp2VHJeNDfnf8ZLh77lX1Ca7imY8hhLM8gwh2QVecObJ-aqxl2yhVUtY88dXQIomCxVCxKjMIcNQl06UBPT-SUE7JCtZtb0MT0YSm6qrA9kVBX8Tsg-5Qgn-lzjwD_QgqSD6nKv58AnH6IGwqsg3Eh5elXVtU-FNRV--iGO5vVVb9kfbjg6JQcd6oLZpXDn3NPuFLCr7n71m1GNmbSIsu0QOZybK11VqMQ0a04VzR2C-kMnGyo2ekvgghAVUw

mWqJJghRaVJSfvv01pDIAbdqk_ARv6GGiAVZ79qBNxK7tLjwrgtsuzpNRfkeMYm3Ta2g
j7jmtgS8EeM22n5cuxauDqO87LgZOTE3asOt64tcUeKNxLIBEsJI5ftARWRwe4VSVXwh
ayfEm_9U9E4bO7m9suHLTILd_6xX2PlRpfHJHynLU7tzI6JE9MVEe6mcI1eAcnGTI0VO
4GmjxC9yN-9R1A1gksIC_Hj7peW376po1ciC3xTxKVyN9KbjdUX0U5e4IBh8CM17ysZ
pZmK1FTpZMAiTCwZRAiyF_Ok73SmlMEi6DMR2HkqT82oQMeu0UUzTmXwbfKdSoeHHbc0
V9I6Rpt7W1b3yz7SZV-JNRvLmXe5b1Kw-kNHwajujUTCGANusLXbXUCEBecxt8JiAj
wV76Ol3Lr1V3p0EoOWAe23vzspgmAJmGqPAM13f5M8T4Ul0diBBrpyUEucJJ4yddtA7L
3itUBLZ8memzQz2za8ZffsbOhR_zZWgrzt-J6-C2ixd0NLgbq88xzDdwpXhTQP-b3c
yeFkqQA_nsGBVduYO9Wh3oVEu3r5Zo-UWLh0ndD4nFdDf0028yIhyF1mFZoucoY84JEs
CzspFe8th9rgFs8m_KxMCdVEQGmg6FNr2z9wmPomZghReJLmdvnVJY7Ee-5wQ0fE_qO
kfh03PRs. 258ci Times Cited:794 Cited References Count:46.

Komivi Dossa, Rong Zhou, Donghua Li, Aili Liu, Lu Qin, Marie A. Mmadi, Ruqi Su, Yujuan Zhang, Jianqiang Wang, Yuan Gao, Xiurong Zhang, and Jun You. A novel motif in the 5'-utr of an orphan gene '¡i¿big root biomass¡/i¿' modulates root biomass in sesame. *Plant Biotechnology Journal*, 19(5):1065–1079, 2021. ISSN 1467-7644. doi: 10.1111/pbi.13531. URL https://dx.doi.org/10.1111/pbi.13531.

R. L. Fernando and M. Grossman. Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution*, 21(4):467–477, 1989. ISSN 0999-193x. doi: DOI10.1051/gse:19890407. URL <GotoISI>://WOS:A1989CK31500007https://gsejournal.biomedcentral.com/counter/pdf/10.1186/1297-9686-21-4-467.pdf. Ck315 Times Cited:383 Cited References Count:14.

A. M. Glazier, J. H. Nadeau, and T. J. Aitman. Finding genes that underlie complex traits. *Science*, 298(5602):2345–9, 2002. ISSN 0036-8075. doi: 10.1126/science.

1076641. URL https://www.science.org/doi/10.1126/science.1076641. 1095-9203 Glazier, Anne M Nadeau, Joseph H Aitman, Timothy J Journal Article Research Support, Non-U.S. Gov't Review United States 2002/12/21 Science. 2002 Dec 20;298(5602):2345-9. doi: 10.1126/science.1076641.

Z. G. Guo, M. M. Magwire, C. J. Basten, Z. Y. Xu, and D. L. Wang. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics*, 129(12):2413–2427, 2016. ISSN 0040-5752. doi: 10.1007/s00122-016-2780-5. URL <GotoISI>://WOS:000389254700016https://link.springer.com/content/pdf/10.1007/s00122-016-2780-5.pdf. Ee0hj Times Cited:54 Cited References Count:55.

A. L. Harfouche, D. A. Jacobson, D. Kainer, J. C. Romero, A. H. Harfouche, G. S. Mugnozza, M. Moshelion, G. A. Tuskan, J. J. B. Keurentjes, and A. Altman. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends in Biotechnology*, 37(11):1217–1235, 2019. ISSN 0167-7799. doi: 10.1016/j.tibtech.2019.05.007. URL <GotoISI>://WOS:000493285300011. Ji2hs Times Cited:87 Cited References Count:100.

E. L. Heffner, M. E. Sorrells, and J. L. Jannink. Genomic selection for crop improvement. *Crop Science*, 49(1):1–12, 2009. ISSN 0011-183x. doi: 10.2135/cropsci2008.08.0512. URL <GotoISI>://WOS:000263344300001. 407dr Times Cited:967 Cited References Count:90.

H. Hu, M. T. Campbell, T. H. Yeats, X. Zheng, D. E. Runcie, G. Covarrubias-Pazaran, C. Broeckling, L. Yao, M. Caffe-Treml, L. A. Gutierrez, K. P. Smith, J. Tanaka, O. A. Hoekenga, M. E. Sorrells, M. A. Gore, and J. L. Jannink. Multi-

omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. *Theor Appl Genet*, 134(12):4043–4054, 2021. ISSN 1432-2242 (Electronic) 0040-5752 (Print) 0040-5752 (Linking). doi: 10.1007/s00122-021-03946-4. URL `https://www.ncbi.nlm.nih.gov/pubmed/34643760https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8580906/pdf/122_2021_Article_3946.pdf`. Hu, Haixiao Campbell, Malachy T Yeats, Trevor H Zheng, Xuying Runcie, Daniel E Covarrubias-Pazaran, Giovanny Broeckling, Corey Yao, Linxing Caffe-Treml, Melanie Gutierrez, Luci A Smith, Kevin P Tanaka, James Hoekenga, Owen A Sorrells, Mark E Gore, Michael A Jannink, Jean-Luc eng 2017-67007-26502/USDA-NIFA-AFRI/ Germany 2021/10/14 Theor Appl Genet. 2021 Dec;134(12):4043-4054. doi: 10.1007/s00122-021-03946-4. Epub 2021 Oct 13.

X. Hu, B. F. Carver, Y. A. El-Kassaby, L. Zhu, and C. Chen. Weighted kernels improve multi-environment genomic prediction. *Heredity (Edinb)*, 130(2):82–91, 2023. ISSN 1365-2540 (Electronic) 0018-067X (Print) 0018-067X (Linking). doi: 10.1038/s41437-022-00582-6. URL `https://www.ncbi.nlm.nih.gov/pubmed/36522412https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9905581/pdf/41437_2022_Article_582.pdf`. Hu, Xiaowei Carver, Brett F El-Kassaby, Yousry A Zhu, Lan Chen, Charles eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2022/12/16 Heredity (Edinb). 2023 Feb;130(2):82-91. doi: 10.1038/s41437-022-00582-6. Epub 2022 Dec 15.

J. Isidro, J. L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M. E. Sorrells. Training set optimization under population structure in genomic selection. *Theor Appl Genet*, 128(1):145–58, 2015. ISSN 1432-2242 (Electronic) 0040-5752 (Print) 0040-5752 (Linking). doi: 10.1007/s00122-014-2418-4. URL

https://www.ncbi.nlm.nih.gov/pubmed/25367380https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4282691/pdf/122_2014_Article_2418.pdf. Isidro, Julio Jannink, Jean-Luc Akdemir, Deniz Poland, Jesse Heslot, Nicolas Sorrells, Mark E eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Germany 2014/11/05 Theor Appl Genet. 2015 Jan;128(1):145-58. doi: 10.1007/s00122-014-2418-4. Epub 2014 Nov 1.

Y. Jia and J. L. Jannink. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192(4):1513–22, 2012. ISSN 1943-2631 (Electronic) 0016-6731 (Print) 0016-6731 (Linking). doi: 10.1534/genetics.112.144246. URL https://www.ncbi.nlm.nih.gov/pubmed/23086217https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3512156/pdf/1513.pdf. Jia, Yi Jannink, Jean-Luc eng Research Support, U.S. Gov't, Non-P.H.S. 2012/10/23 Genetics. 2012 Dec;192(4):1513-22. doi: 10.1534/genetics.112.144246. Epub 2012 Oct 19.

Kayla A. Johnson and Arjun Krishnan. Robust normalization and transformation techniques for constructing gene coexpression networks from rna-seq data. *Genome Biology*, 23(1), 2022. ISSN 1474-760X. doi: 10.1186/s13059-021-02568-9. URL https://dx.doi.org/10.1186/s13059-021-02568-9.

D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat Biotechnol*, 37(8): 907–915, 2019. ISSN 1546-1696 (Electronic) 1087-0156 (Print) 1087-0156 (Linking). doi: 10.1038/s41587-019-0201-4. URL https://www.ncbi.nlm.nih.gov/pubmed/31375807https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7605509/pdf/nihms-1635086.pdf. Kim, Daehwan Paggi, Joseph M Park, Chanhee Bennett, Christopher Salzberg, Steven L eng R01 HG006102/HG/NHGRI NIH HHS/

R01 HG006677/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2019/08/04 Nat Biotechnol. 2019 Aug;37(8):907-915. doi: 10.1038/s41587-019-0201-4. Epub 2019 Aug 2.

Anna V. Klepikova, Artem S. Kasianov, Evgeny S. Gerasimov, Maria D. Logacheva, and Aleksey A. Penin. A high resolution map of the ¡i¿arabidopsis thaliana¡/i¿ developmental transcriptome based on ¡scp¿rna¡/scp¿-seq profiling. *The Plant Journal*, 88(6):1058–1070, 2016. ISSN 0960-7412. doi: 10.1111/tpj.13312. URL https://dx.doi.org/10.1111/tpj.13312.

R. Lande and R. Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–56, 1990. ISSN 0016-6731 (Print) 0016-6731. doi: 10.1093/genetics/124.3.743. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1203965/pdf/ge1243743.pdf. Lande, R Thompson, R GM27120/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United States 1990/03/01 Genetics. 1990 Mar;124(3):743-56. doi: 10.1093/genetics/124.3.743.

B. Lenaerts, B. C. Y. Collard, and M. Demont. Review: Improving global food security through accelerated plant breeding. *Plant Science*, 287, 2019. ISSN 0168-9452. doi: ARTN11020710.1016/j.plantsci.2019.110207. URL <GotoISI>://WOS:000487165500002https://www.sciencedirect.com/science/article/pii/S0168945219304819?via%3Dihub. Iz6az Times Cited:103 Cited References Count:98.

D. Li, Z. Xu, R. Gu, P. Wang, D. Lyle, J. Xu, H. Zhang, and G. Wang. Enhancing genomic selection by fitting large-effect snps as fixed effects and a genotype-by-environment effect using a maize bc1f3:4 population. *PLoS One*, 14(10):e0223898,

2019. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0223898. URL https://www.ncbi.nlm.nih.gov/pubmed/31622400https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6797203/pdf/pone.0223898.pdf. Li, Dongdong Xu, Zhenxiang Gu, Riliang Wang, Pingxi Lyle, Demar Xu, Jialiang Zhang, Hongwei Wang, Guogying eng Research Support, Non-U.S. Gov't 2019/10/18 PLoS One. 2019 Oct 17;14(10):e0223898. doi: 10.1371/journal.pone.0223898. eCollection 2019.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Print) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btp352. URL https://www.ncbi.nlm.nih.gov/pubmed/19505943https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/pdf/btp352.pdf. Li, Heng Handsaker, Bob Wysoker, Alec Fennell, Tim Ruan, Jue Homer, Nils Marth, Gabor Abecasis, Goncalo Durbin, Richard eng U54HG002750/HG/NHGRI NIH HHS/ R01 HG004719-01/HG/NHGRI NIH HHS/ U54 HG002750/HG/NHGRI NIH HHS/ 077192/Z/05/Z/WT_/Wellcome Trust/United Kingdom R01 HG004719-02/HG/NHGRI NIH HHS/ R01 HG004719/HG/NHGRI NIH HHS/ R01 HG004719-04/HG/NHGRI NIH HHS/ R01 HG004719-02S1/HG/NHGRI NIH HHS/ R01 HG004719-03/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2009/06/10 Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.

Yongle Li, Pradeep Ruperao, Jacqueline Batley, David Edwards, William Martin,

Kristy Hobson, and Tim Sutton. Genomic prediction of preliminary yield trials in chickpea: Effect of functional annotation of snps and environment. *The Plant Genome*, 15(1), 2022. ISSN 1940-3372. doi: 10.1002/tpg2.20166. URL https://dx.doi.org/10.1002/tpg2.20166.

A. J. Lorenz, K. P. Smith, and J. -L Jannink. Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Science*, 52 (4):1609–1621, 2012. ISSN 0011-183X. doi: 10.2135/cropsci2011.09.0503. URL https://dx.doi.org/10.2135/cropsci2011.09.0503.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), 2014. ISSN 1474-760x. doi: ARTN55010.1186/s13059-014-0550-8. URL <GotoISI>://WOS:000346609500022https://genomebiology.biomedcentral.com/counter/pdf/10.1186/s13059-014-0550-8.pdf. Aw9xx Times Cited:36588 Cited References Count:64.

D. H. Lyra, L. D. Mendonça, G. Galli, F. C. Alves, I. S. C. Granato, and R. Fritsche-Neto. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Molecular Breeding*, 37(6), 2017. ISSN 1380-3743. doi: ARTN8010.1007/s11032-017-0681-1. URL <GotoISI>://WOS:000404210900010https://link.springer.com/content/pdf/10.1007/s11032-017-0681-1.pdf. Ey7za Times Cited:35 Cited References Count:66.

T. F. Mackay. The genetic architecture of quantitative traits. *Annu Rev Genet*, 35: 303–39, 2001. ISSN 0066-4197 (Print) 0066-4197. doi: 10.1146/annurev.genet.35.102401.090633. Mackay, T F Journal Article Research Support, U.S. Gov't,

P.H.S. Review United States 2001/11/09 Annu Rev Genet. 2001;35:303-39. doi: 10.1146/annurev.genet.35.102401.090633.

Martin Mascher. Pseudomolecules and annotation of the second version of the reference genome sequence assembly of barley cv. morex [morex v2], 2019. URL https://doi.ipk-gatersleben.de:443/DOI/83e8e186-dc4b-47f7-a820-28ad37cb176b/d1067eba-1d08-42e2-85ec-66bfd5112cd8/2.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001. ISSN 0016-6731. URL <GotoISI>://WOS:000168223400036https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461589/pdf/11290733.pdf. 424gd Times Cited:4807 Cited References Count:22.

S. Michel, C. Wagner, T. Nosenko, B. Steiner, M. Samad-Zamini, M. Buerstmayr, K. Mayer, and H. Buerstmayr. Merging genomics and transcriptomics for predicting fusarium head blight resistance in wheat. *Genes*, 12(1), 2021. doi: ARTN11410.3390/genes12010114. URL <GotoISI>://WOS:000610262000001https://mdpi-res.com/d_attachment/genes/genes-12-00114/article_deploy/genes-12-00114.pdf?version=1611030435. Pv8wq Times Cited:8 Cited References Count:57.

O. A. Montesinos-Lopez, J. C. Montesinos-Lopez, E. Salazar, J. A. Barron, A. Montesinos-Lopez, R. Buenrostro-Mariscal, and J. Crossa. Application of a poisson deep neural network model for the prediction of count data in genome-based prediction. *Plant Genome*, 14(3), 2021. doi: ARTNe2011810.1002/tpg2.20118.

URL <GotoISI>://WOS:000678659900001. Yy3xc Times Cited:6 Cited References Count:38.

Masatoshi Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973. ISSN 0027-8424. doi: 10.1073/pnas.70.12.3321. URL https://dx.doi.org/10.1073/pnas.70.12.3321 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC427228/pdf/pnas00139-0051.pdf.

H. P. Piepho and J. Möhring. Computing heritability and selection response from unbalanced plant breeding trials. *Genetics*, 177(3):1881–1888, 2007. ISSN 0016-6731. doi: 10.1534/genetics.107.074229. URL <GotoISI>://WOS:000251368800050https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2147938/pdf/GEN1773 1881.pdf. 237ja Times Cited:389 Cited References Count:21.

M. Pourkheirandish and T. Komatsuda. The importance of barley genetics and domestication in a global perspective. *Annals of Botany*, 100(5):999–1008, 2007. ISSN 0305-7364. doi: 10.1093/aob/mcm139. URL <GotoISI>://WOS:000250663800010https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2759206/pdf/mcm139.pdf. 227nq Times Cited:103 Cited References Count:103.

M. D. Purugganan and D. Q. Fuller. The nature of selection during plant domestication. *Nature*, 457(7231):843–848, 2009. ISSN 0028-0836. doi: 10.1038/nature07895. URL <GotoISI>://WOS:000263266700035https://www.nature.com/articles/nature07895.pdf. 406af Times Cited:638 Cited References Count:74.

R. Rincent, D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C. C. Schoen, N. Meyer, C. Giauf-

fret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (l.). *Genetics*, 192(2):715–+, 2012. ISSN 0016-6731. doi: 10.1534/genetics.112.141473. URL <GotoISI>://WOS:000309547400026https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3454892/pdf/715.pdf. 016vi Times Cited:206 Cited References Count:50.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3), 2010. ISSN 1474-760x. doi: ARTNR2510.1186/gb-2010-11-3-r25. URL <GotoISI>://WOS:000277309100013https://genomebiology.biomedcentral.com/counter/pdf/10.1186/gb-2010-11-3-r25.pdf. 591ny Times Cited:4579 Cited References Count:31.

J. Rutkoski, J. Poland, S. Mondal, E. Autrique, L. G. Pérez, J. Crossa, M. Reynolds, and R. Singh. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3-Genes Genomes Genetics*, 6(9):2799–2808, 2016. ISSN 2160-1836. doi: 10.1534/g3.116.032888. URL <GotoISI>://WOS:000384021400012https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015937/pdf/2799.pdf. Dw9zn Times Cited:224 Cited References Count:23.

K. Sandhu, S. S. Patil, M. Pumphrey, and A. Carter. Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. *Plant Genome*, 14(3), 2021. doi: ARTNe2011910.1002/tpg2.2

0119. URL <GotoISI>://WOS:000692501000001. Yy3xc Times Cited:42 Cited References Count:68.

T. A. Schrag, M. Westhues, W. Schipprack, F. Seifert, A. Thiemann, S. Scholten, and A. E. Melchinger. Beyond genomic prediction: Combining different types of data can improve prediction of hybrid performance in maize. *Genetics*, 208(4): 1373–1385, 2018. ISSN 0016-6731. doi: 10.1534/genetics.117.300374. URL <GotoISI>://WOS:000429094400007https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5887136/pdf/1373.pdf. Gb5hf Times Cited:90 Cited References Count:67.

N. J. Schurch, P. Schofield, M. Gierlinski, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton. How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–51, 2016. ISSN 1469-9001 (Electronic) 1355-8382 (Print) 1355-8382 (Linking). doi: 10.1261/rna.053959.115. URL https://www.ncbi.nlm.nih.gov/pubmed/27022035. Schurch, Nicholas J Schofield, Pieta Gierlinski, Marek Cole, Christian Sherstnev, Alexander Singh, Vijender Wrobel, Nicola Gharbi, Karim Simpson, Gordon G Owen-Hughes, Tom Blaxter, Mark Barton, Geoffrey J eng WT097945/WT_/Wellcome Trust/United Kingdom 095062/WT_/Wellcome Trust/United Kingdom WT083481/WT_/Wellcome Trust/United Kingdom BB/H002286/1/BB_/Biotechnology and Biological Sciences Research Council/United Kingdom BB/M004155/1/BB_/Biotechnology and Biological Sciences Research Council/United Kingdom 92530/Z/10/Z/WT_/Wellcome Trust/United Kingdom MR/K001744/1/MRC_/Medical Research Council/United Kingdom

WT092340/WT_/Wellcome Trust/United Kingdom 098439/Z/12/WT_/Wellcome Trust/United Kingdom 2016/03/30 RNA. 2016 Jun;22(6):839-51. doi: 10.1261/rna.053959.115. Epub 2016 Mar 28.

S. R. Searle, F. M. Speed, and G. A. Milliken. Population marginal means in the linear-model - an alternative to least-squares means. *American Statistician*, 34(4): 216–221, 1980. ISSN 0003-1305. doi: Doi10.2307/2684063. URL <GotoISI>://WOS:A1980KT98100004. Kt981 Times Cited:955 Cited References Count:7.

A. Shrestha, F. Cosenza, D. van Inghelandt, P. Y. Wu, J. Q. Li, F. A. Casale, M. Weisweiler, and B. Stich. The double round-robin population unravels the genetic architecture of grain size in barley. *Journal of Experimental Botany*, 73(22): 7344–7361, 2022. ISSN 0022-0957. doi: 10.1093/jxb/erac369. URL <GotoISI>://WOS:000878756900001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9730814/pdf/erac369.pdf. 6w5vu Times Cited:3 Cited References Count:85.

B. Stich. Comparison of mating designs for establishing nested association mapping populations in maize and. *Genetics*, 183(4):1525–1534, 2009. ISSN 0016-6731. doi: 10.1534/genetics.109.108449. URL <GotoISI>://WOS:000272435000025https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2787436/pdf/GEN18341525.pdf. 528hv Times Cited:36 Cited References Count:43.

Biyue Tan, Dario Grattapaglia, Gustavo Salgado Martins, Karina Zamprogno Ferreira, Björn Sundberg, and Pär K. Ingvarsson. Evaluating the accuracy of genomic prediction of growth and wood traits in two eucalyptus species and their f1 hybrids. *BMC Plant Biology*, 17(1), 2017. ISSN 1471-2229. doi: 10.1186/s12870-017-1059-6. URL https://dx.doi.org/10.1186/s12870-017-1059-6.

D. Tilman, C. Balzer, J. Hill, and B. L. Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20260–20264, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1116437108. URL <GotoISI>://WOS:000298034800082https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3250154/pdf/pnas.201116437.pdf. 861qr Times Cited:4414 Cited References Count:31.

H. Y. Tsai, F. Cericola, V. Edriss, J. R. Andersen, J. Orabiid, J. D. Jensen, A. Jahoor, L. Janss, and J. Jensen. Use of multiple traits genomic prediction, genotype by environment interactions and spatial effect to improve prediction accuracy in yield data. *Plos One*, 15(5), 2020. ISSN 1932-6203. doi: ARTNe023266510.1371/journal.pone.0232665. URL <GotoISI>://WOS:000537481000029https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0232665&type=printable. Lu0vg Times Cited:17 Cited References Count:44.

Jordan Ubbens, Isobel Parkin, Christina Eynck, Ian Stavness, and Andrew G. Sharpe. Deep neural networks for genomic prediction do not estimate marker effects. *The Plant Genome*, 14(3):e20147, 2021. ISSN 1940-3372. doi: https://doi.org/10.1002/tpg2.20147. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/tpg2.20147.

P. M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008. ISSN 0022-0302. doi: 10.3168/jds.2007-0980. URL <GotoISI>://WOS:000260277200035https://www.sciencedirect.com/science/article/pii/S0022030208709901?via%3Dihub. 363os Times Cited:3552 Cited References Count:24.

J. D. Washburn, M. B. Burch, and J. A. V. Franco. Predictive breeding for maize: Making use of molecular phenotypes, machine learning, and physiological crop models. *Crop Science*, 60(2):622–638, 2020. ISSN 0011-183x. doi: 10.1002/csc2.20052. URL <GotoISI>://WOS:000540510200010. Ly4oz Times Cited:26 Cited References Count:146.

Sven E. Weber, Lennard Roscher-Ehrig, Tobias Kox, Amine Abbadi, Andreas Stahl, and Rod J. Snowdon. Genomic prediction in ¡i¿brassica napus¡/i¿: evaluating the benefit of imputed whole-genome sequencing data. *Genome*, 67(7):210–222, 2024. ISSN 0831-2796. doi: 10.1139/gen-2023-0126. URL https://dx.doi.org/10.1139/gen-2023-0126.

M. Weisweiler, A. de Montaigu, D. Ries, M. Pfeifer, and B. Stich. Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mrna sequencing and their power to predict phenotypic traits. *Bmc Genomics*, 20(1), 2019. ISSN 1471-2164. doi: ARTN78710.1186/s12864-019-6174-3. URL <GotoISI>://WOS:000500823900005https://bmcgenomics.biomedcentral.com/counter/pdf/10.1186/s12864-019-6174-3.pdf. Jt2ke Times Cited:14 Cited References Count:59.

M. Weisweiler, C. Arlt, P. Y. Wu, D. Van Inghelandt, T. Hartwig, and B. Stich. Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. *Theoretical and Applied Genetics*, 135(10):3511–3529, 2022. ISSN 0040-5752. doi: 10.1007/s00122-022-04197-7. URL <GotoISI>://WOS:000846305700001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC95196

79/pdf/122_2022_Article_4197.pdf. 5a1yz Times Cited:4 Cited References Count:78.

Z. X. Wen, R. J. Tan, S. C. Zhang, P. J. Collins, J. Z. Yuan, W. Y. Du, C. H. Gu, S. J. Ou, Q. J. Song, Y. Q. C. An, J. F. Boyse, M. I. Chilvers, and D. C. Wang. Integrating gwas and gene expression data for functional characterization of resistance to white mould in soya bean. *Plant Biotechnology Journal*, 16(11): 1825–1835, 2018. ISSN 1467-7644. doi: 10.1111/pbi.12918. URL <GotoISI>://WOS:000446996000001https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6181214/pdf/PBI-16-1825.pdf. Gw5rz Times Cited:42 Cited References Count:55.

M. Westhues, T. A. Schrag, C. Heuer, G. Thaller, H. F. Utz, W. Schipprack, A. Thiemann, F. Seifert, A. Ehret, A. Schlereth, M. Stitt, Z. Nikoloski, L. Willmitzer, C. C. Schön, S. Scholten, and A. E. Melchinger. Omics-based hybrid prediction in maize. *Theoretical and Applied Genetics*, 130(9):1927–1939, 2017. ISSN 0040-5752. doi: 10.1007/s00122-017-2934-0. URL <GotoISI>://WOS:000408120800012https://link.springer.com/content/pdf/10.1007/s00122-017-2934-0.pdf. Fe3mu Times Cited:71 Cited References Count:90.

M. Westhues, C. Heuer, G. Thaller, R. Fernando, and A. E. Melchinger. Efficient genetic value prediction using incomplete omics data. *Theoretical and Applied Genetics*, 132(4):1211–1222, 2019. ISSN 0040-5752. doi: 10.1007/s00122-018-03273-1. URL <GotoISI>://WOS:000463674000026https://link.springer.com/content/pdf/10.1007/s00122-018-03273-1.pdf. Hs2en Times Cited:6 Cited References Count:55.

S. I. Wright. The effects of artificial selection on the maize genome (vol 308, pg 1310, 2005). *Science*, 310(5745):54–54, 2005. ISSN 0036-8075. URL <GotoISI>://WOS:000232477000025. 972tv Times Cited:3 Cited References Count:1.

P. Y. Wu, B. Stich, M. Weisweiler, A. Shrestha, A. Erban, P. Westhoff, and D. Van Inghelandt. Improvement of prediction ability by integrating multi-omic datasets in barley. *Bmc Genomics*, 23(1), 2022. ISSN 1471-2164. doi: ARTN20010.1186/s12864-022-08337-7. URL <GotoISI>://WOS:000767937100003https://bmcgenomics.biomedcentral.com/counter/pdf/10.1186/s12864-022-08337-7.pdf. Zr7cu Times Cited:5 Cited References Count:52.

F. Zabel, R. Delzeit, J. M. Schneider, R. Seppelt, W. Mauser, and T. Václavík. Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity. *Nature Communications*, 10, 2019. doi: ARTN284410.1038/s41467-019-10775-z. URL <GotoISI>://WOS:000473132200008https://www.nature.com/articles/s41467-019-10775-z.pdf. If5on Times Cited:287 Cited References Count:75.

K. Zarringhalam, D. Degras, C. Brockel, and D. Ziemek. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. *Sci Rep*, 8(1):1237, 2018. ISSN 2045-2322 (Electronic) 2045-2322 (Linking). doi: 10.1038/s41598-018-19635-0. URL https://www.ncbi.nlm.nih.gov/pubmed/29352257. Zarringhalam, Kourosh Degras, David Brockel, Christoph Ziemek, Daniel eng R01 AI167570/AI/NIAID NIH HHS/ R21 AI150090/AI/NIAID NIH HHS/ U54 CA156734/CA/NCI NIH HHS/ England 2018/01/21 Sci Rep. 2018 Jan 19;8(1):1237. doi: 10.1038/s41598-018-19635-0.

Chaozhong Zhang, Lin Huang, Huifei Zhang, Qunqun Hao, Bo Lyu, Meinan Wang, Lynn Epstein, Miao Liu, Chunlan Kou, Juan Qi, Fengjuan Chen, Mengkai Li, Ge Gao, Fei Ni, Lianquan Zhang, Ming Hao, Jirui Wang, Xianming Chen, Ming-Cheng Luo, Youliang Zheng, Jiajie Wu, Dengcai Liu, and Daolin Fu. An ancestral nb-lrr with duplicated 3'utrs confers stripe rust resistance in wheat and barley. *Nature Communications*, 10(1), 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11872-9. URL https://dx.doi.org/10.1038/s41467-019-11872-9.

Y. Zhao, M. F. Mette, M. Gowda, C. F. H. Longin, and J. C. Reif. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity*, 112(6):638–645, 2014. ISSN 0018-067x. doi: 10.1038/hdy.2014.1. URL <GotoISI>://WOS:000336501000009https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4023446/pdf/hdy20141a.pdf. Ah9ye Times Cited:114 Cited References Count:45.

S. Q. Zhong, J. C. M. Dekkers, R. L. Fernando, and J. L. Jannink. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, 182(1):355–364, 2009. ISSN 0016-6731. doi: 10.1534/genetics.108.098277. URL <GotoISI>://WOS:000270213800030https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674832/pdf/GEN1821355.pdf. 499ix Times Cited:286 Cited References Count:25.

X. T. Zhu, H. P. Maurer, M. Jenz, V. Hahn, A. Ruckelshausen, W. L. Leiser, and T. Würschum. The performance of phenomic selection depends on the genetic architecture of the target trait. *Theoretical and Applied Genetics*, 135(2):653–665, 2022. ISSN 0040-5752. doi: 10.1007/s00122-021-03997-7. URL <GotoISI>://WOS:000721435300002https://www.ncbi.nlm.nih.gov/pmc/articles/PMC88663

87/pdf/122_2021_Article_3997.pdf. Zg6gq Times Cited:7 Cited References Count:48.

STATEMENTS & DECLARATIONS

Funding

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID:390686111).

Competing Interests

Benjamin Stich is a member of the editorial board of *Theoretical and Applied Genetics*. Furthermore, the authors have no competing interests to declare that are relevant to the content of this article.

Author contributions

The study was conceptualized and designed by C.A and B.S.; The RIL population was designed by B.S and J.L; The genetic material was propagated and maintained by D.I and B.S; The data were analyzed and interpreted by C.A and B.S; The manuscript was written by C.A. and edited by C.A and B.S; All authors read and approved the final manuscript.

Data availability

The raw read RNA-Seq dataset analyzed in this study is available in the NCBI Sequence Read Archive (SRA), BioProject ID: PRJNA1088431, URL:

https://www.ncbi.nlm.nih.gov/bioproject/1088431

https://www.ncbi.nlm.nih.gov/sra/PRJNA1088431

Table 1. Overview of all studied phenotypic traits. Number of environments in which the trait was assessed, broad sense heritability (H^2), minimum (Min.), maximum (Max.), median, mean, and standard deviation (SD) of adjusted entry means of all traits included in this study.

Traits	Environments	H^2	Min.	Max.	Median	Mean	SD
Ear/spike length [cm]	5	0.88	6.04	13.64	8.96	8.92	1.28
Awn length [cm]	5	0.74	10.07	17.60	13.45	13.52	1.39
Flowering time [d]	7	0.86	55.73	87.74	70.36	69.62	6.10
Plant height [cm]	5	0.76	38.90	99.28	69.60	69.67	9.34
Grain length [mm]	4	0.84	7.35	13.92	10.69	10.84	1.02
Grain width [mm]	4	0.85	2.80	3.76	3.41	3.40	0.18
Grain area [mm^2]	4	0.89	16.48	30.83	25.59	25.28	2.66
Thousand grain weight [g]	4	0.84	25.10	55.09	37.76	37.96	5.73

Table 2. Overview of all predictor datasets. Name, abbreviation, origin, and number of features (before filtering) for all genomic and transcriptomic datasets included in this study.

Name	Abbreviation	Origin	No. of features
Total transcript expression	GE^{Total}_{RNAseq}	RNA-Seq	42.6K
Filtered transcript expression	$GE^{Filter.}_{RNAseq}$	RNA-Seq	37.7K
Differentially expressed transcript expression	GE^{DEG}_{RNAseq}	RNA-Seq	7.3K
Unfiltered transcriptome sequence variants	SNP^{Total}_{RNAseq}	RNA-Seq	147.5K
Functional transcriptome sequence regulatory variants	$SNP^{Reg.}_{RNAseq}$	RNA-Seq	81.7K
Functional transcriptome sequence non-synonymous variants	SNP^{CDS}_{RNAseq}	RNA-Seq	25.8K
Standard filtered transcriptome sequence variants	$SNP^{Stn.}_{RNAseq}$	RNA-Seq	52.2K
Quality filtered transcriptome sequence variants	SNP^{QC}_{RNAseq}	RNA-Seq	42.5K
RNA-Seq / WGS consensus sequence variants	SNP^{Total}_{WGS}	Hybrid	426.4K
RNA-Seq / WGS consensus sequence non-synonymous variants	SNP^{CDS}_{WGS}	Hybrid	52.4K
RNA-Seq / WGS consensus sequence regulatory variants	$SNP^{Reg.}_{WGS}$	Hybrid	245.6K
50k SNP array data	SNP^{Total}_{Array}	Array	17.3K

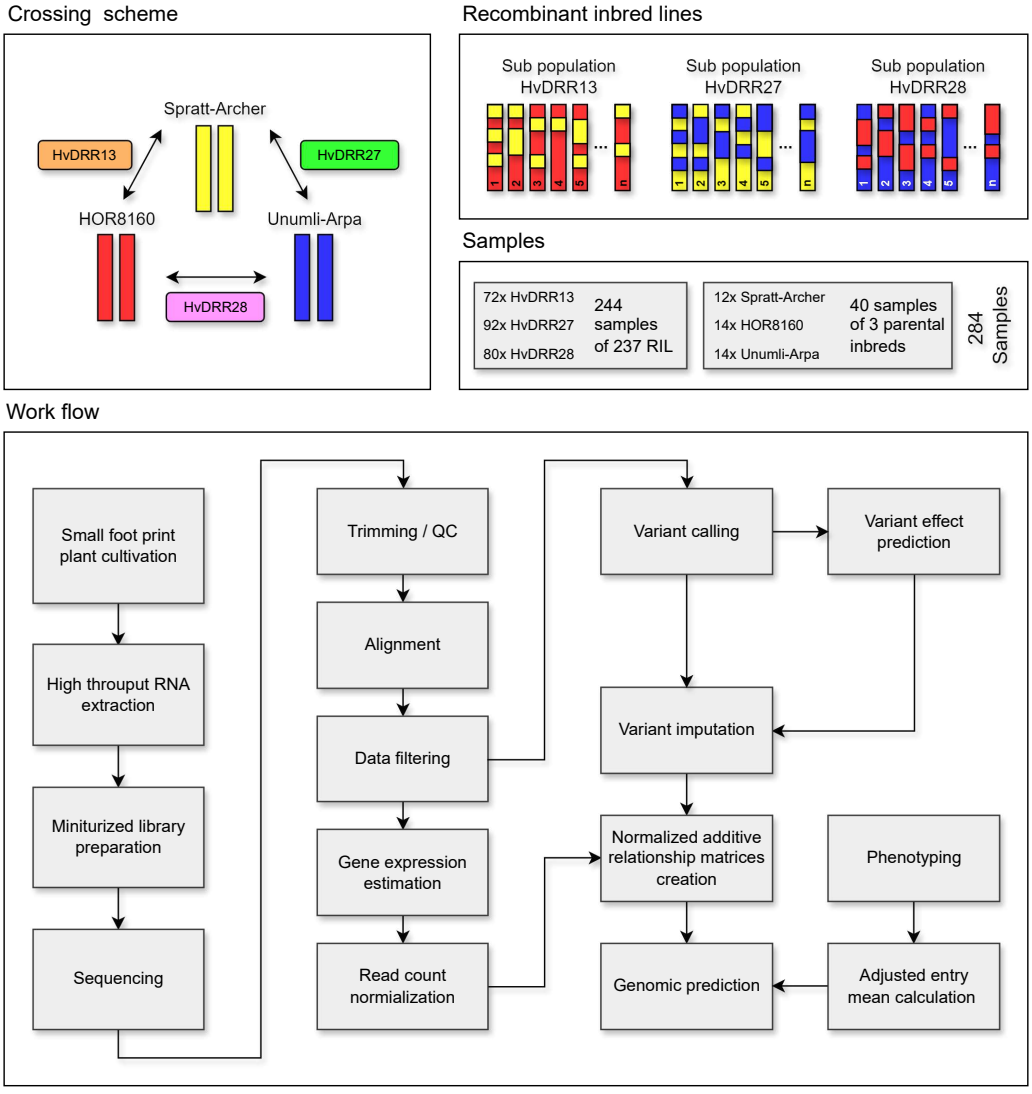


Fig. 1. Genetic material and workflow overview. The crossing scheme shows the three homozygous parental inbreds that were used to create the recombinant inbred line populations HvDRR13, HvDRR27, and HvDRR28 (F7/F8). The workflow is shown by connecting the major steps in consecutive order.

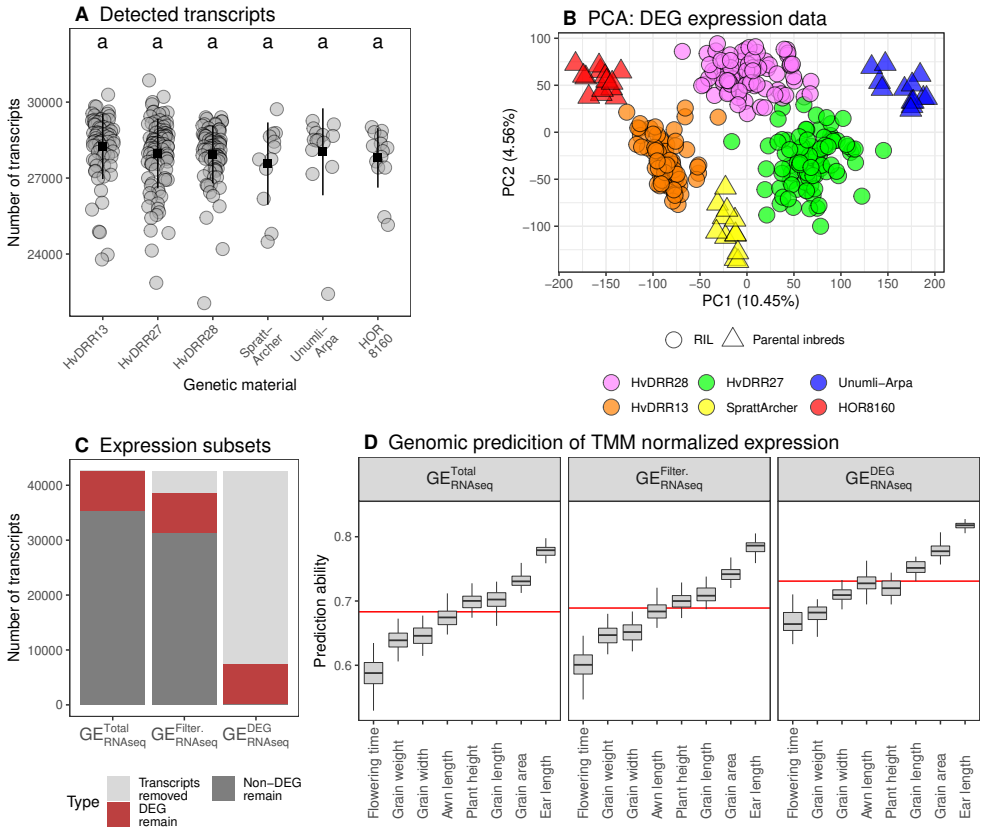


Fig. 2. Characteristics and genomic prediction (GP) performance of the gene expression datasets. (A) The number of detected transcripts are grouped by genetic material with mean (black square) and standard deviation (black line). Populations marked by the same letter are not significantly ($\alpha = 0.05$) different from each other. (B) Principle component analysis using the differentially expressed genes (DEG) subset (GE^{DEG}_{RNAseq}). PC 1 and PC 2 are the first and second principal component, respectively, and the number in parentheses refers to the proportion of variance explained by the principal components in percent. (C) Number of transcripts included in all data subsets used for genomic prediction (GP). (D) GP results are compared between all gene expression subsets and all traits included in this study. The red line shows the mean prediction ability per subset.

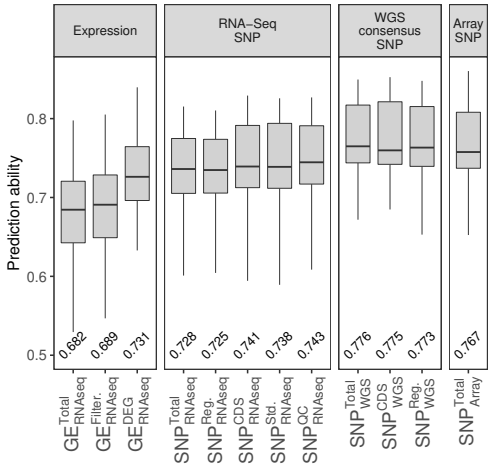


Fig. 3. Prediction ability of four dataset groups: Expression, RNA-Seq SNP, WGS consensus SNP, and SNP array across all the eight examined traits. For each dataset, the mean prediction ability value is shown as number in the bottom.

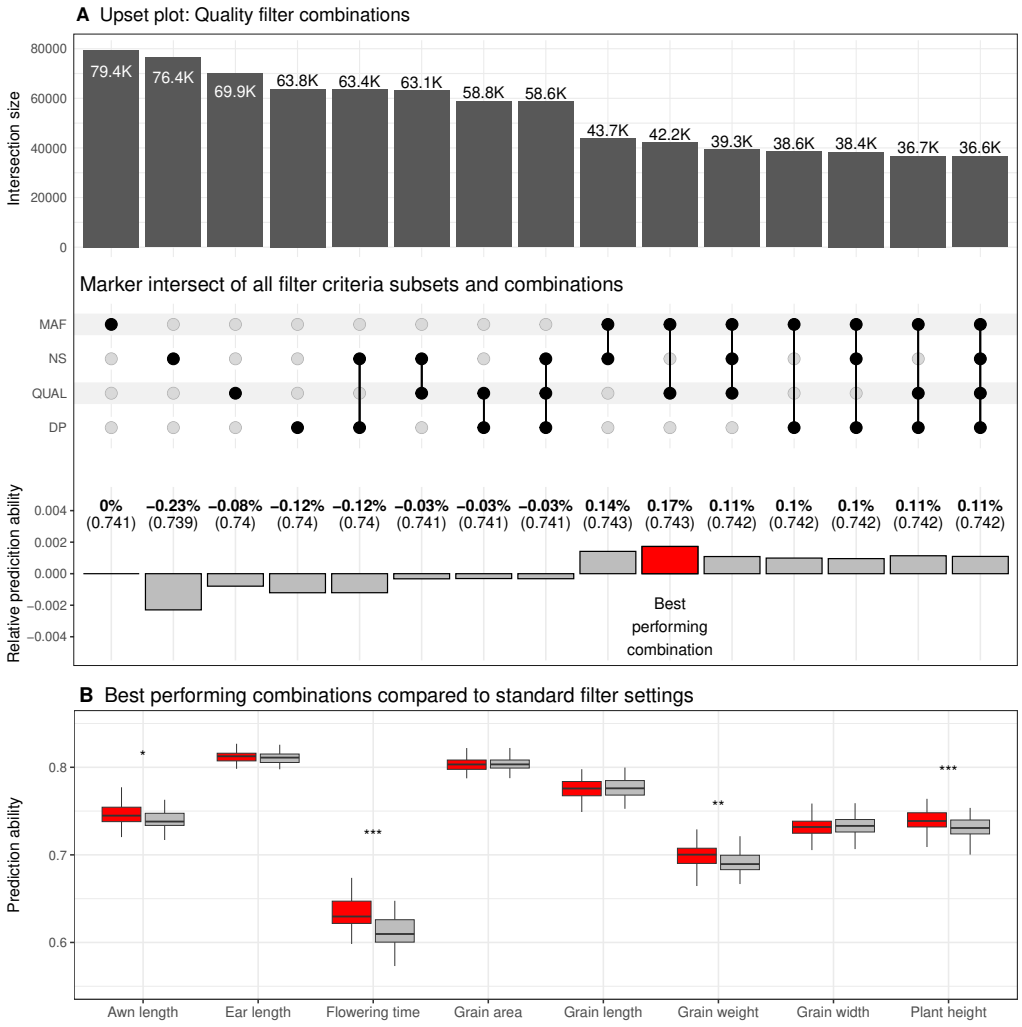


Fig. 4. (A) Combination of quality filtering subsets of genomic variants and their relative genomic prediction (GP) performance. For each of the combinations, the number of remaining variants (top) are shown after selecting the inclusive intersect between all included filtering subsets (middle). The four quality filtering criteria are minor allele frequency (MAF), variant quality score (QUAL), missing rate (NS), and read depth (DP). The subsets are the best performing subsets that were marked in Fig 4 with "T". The relative GP performance (bottom) shows the difference in prediction ability in percent between the combined subsets and the top performing single quality filtered subset (DP). The best performing subset is marked (red). (B) GP performance comparison between a standard quality filtering subset (white) (NA $<$ 20%, MAF $>$ 0.05) and the best performing combination (red). The GP results were averaged across all eight traits. Significant differences: p-value $<$ 0.01 = **, p-value $<$ 0.001 = ***.

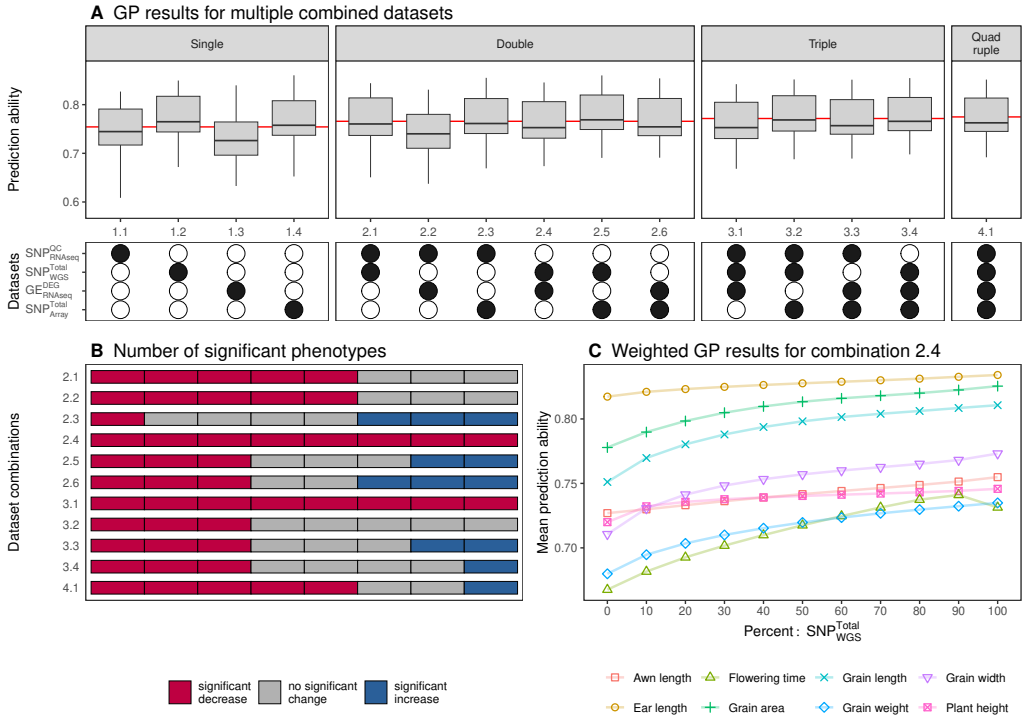


Fig. 5. Overview of genomic prediction abilities using multiple datasets. (A) The four different single datasets: SNP^{QC}_{RNAseq}, SNP^{Total}_{WGS}, GE^{DEG}_{RNAseq}, and SNP^{Total}_{Array} were combined and prediction ability was calculated (top). The combinations are indicated as a combination of white (not included) and black (included) dots (bottom). The red line shows the mean prediction ability within each combination group. (B) The number of significantly ($p > 0.05$) increased (blue), decreased (red), and not significantly changed (grey) phenotypic traits compared to the value of the single best dataset for each combination. (C) The mean prediction ability is plotted for the weighted combination of SNP^{Total}_{WGS} and GE^{DEG}_{RNAseq} (2.4). The weight is gradually changed from 100% GE^{DEG}_{RNAseq} (left) to 100% SNP^{Total}_{WGS} (right) with a step-size of 10%. The results are shown for each weighted combination and all traits.

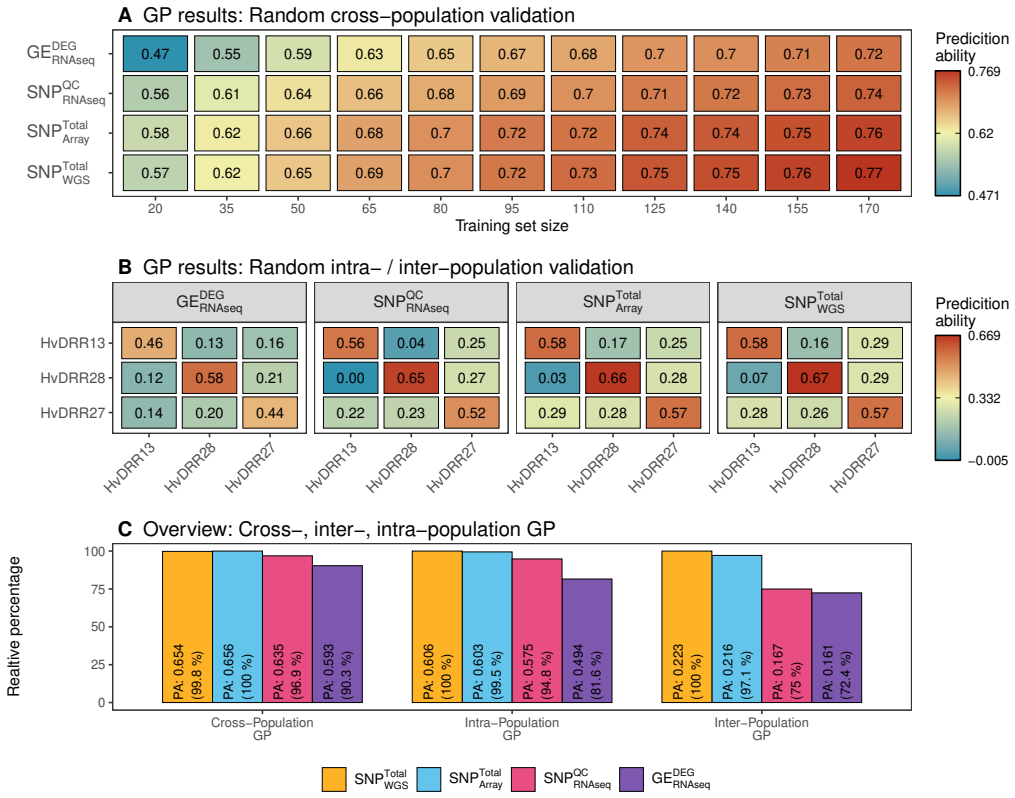


Fig. 6. (A) Genomic prediction (GP) results of the best performing datasets from each data origin group using randomly selected cross-population training sets (TS). The TS size was gradually increased from 20 to 170 individuals (step-size: 15). (B) GP results for intra-population (diagonal) or inter-population (off diagonal) comparisons. (C) Prediction ability (PA) deviation between the four main datasets for cross-, intra-, and inter-population genomic prediction. PA was averaged over all random validation runs and all intra- and inter-population pairwise comparisons. The training sets sizes was set to 50 for (B) and (C).

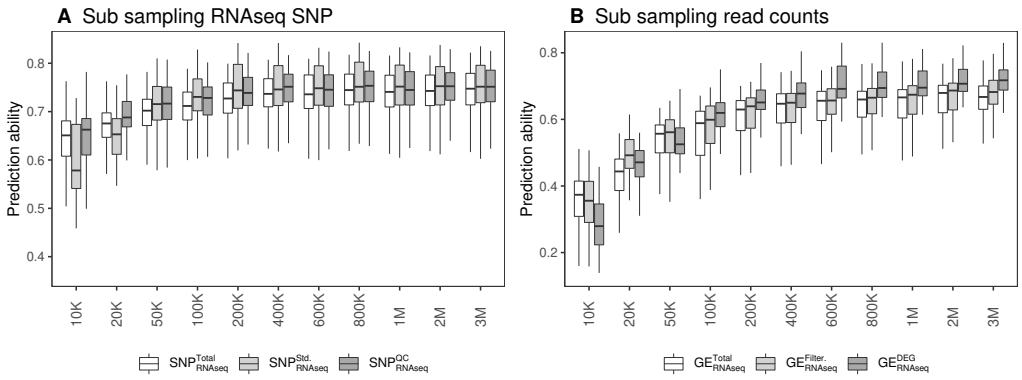


Fig. 7. Genomic prediction (GP) performance of artificially reduced sequencing depth subsets created by random read sub-sampling of (A) RNA-Seq genomic variant datasets (SNP^{Total}_{RNAseq}, SNP^{QC}_{RNAseq}) and (B) gene expression datasets (GE^{Total}_{RNAseq}, $GE^{Filter.}_{RNAseq}$, GE^{DEG}_{RNAseq}) including 229 of the 240 samples. The sub-sampling ranged in 11 steps between 10 thousand to 3 million reads.

Table S1. GP performance changes of all dataset combinations compared to SNP_{Array}^{Total} (here: 1.4). All traits with significantly increased prediction abilities (Up), significantly decreased prediction abilities (Down), and not significantly changed prediction abilities (NC) were counted and named (EL: ear length, AL: awn length, PH: plant height, FT: flowering time, GL: grain length, GW: grain width, GA: grain area, TGW: grain weight). Dataset names are analogous to Fig. 5.

Dataset	Down	Up	NC	Decreased traits	Increased traits
1.1	7	1	0	EL,PH,FT,GL,GW,GA,TGW	AL
1.2	2	5	1	GA,TGW	EL,AL,FT,GL,GW
1.3	7	0	1	AL,PH,FT,GL,GW,GA,TGW	
1.4	0	0	8		
2.1	2	4	2	GA,TGW	EL,AL,FT,GL
2.2	7	1	0	EL,PH,FT,GL,GW,GA,TGW	AL
2.3	1	4	3	GA	EL,AL,FT,GW
2.4	4	4	0	PH,GW,GA,TGW	EL,AL,FT,GL
2.5	1	6	1	GA	EL,AL,FT,GL,GW,TGW
2.6	3	3	2	PH,GA,TGW	EL,AL,FT
3.1	4	4	0	PH,GW,GA,TGW	EL,AL,FT,GL
3.2	1	5	2	GA	EL,AL,FT,GL,GW
3.3	3	3	2	PH,GA,TGW	EL,AL,FT
3.4	1	5	2	GA	EL,AL,FT,GL,GW
4.1	2	5	1	GA,TGW	EL,AL,FT,GL,GW

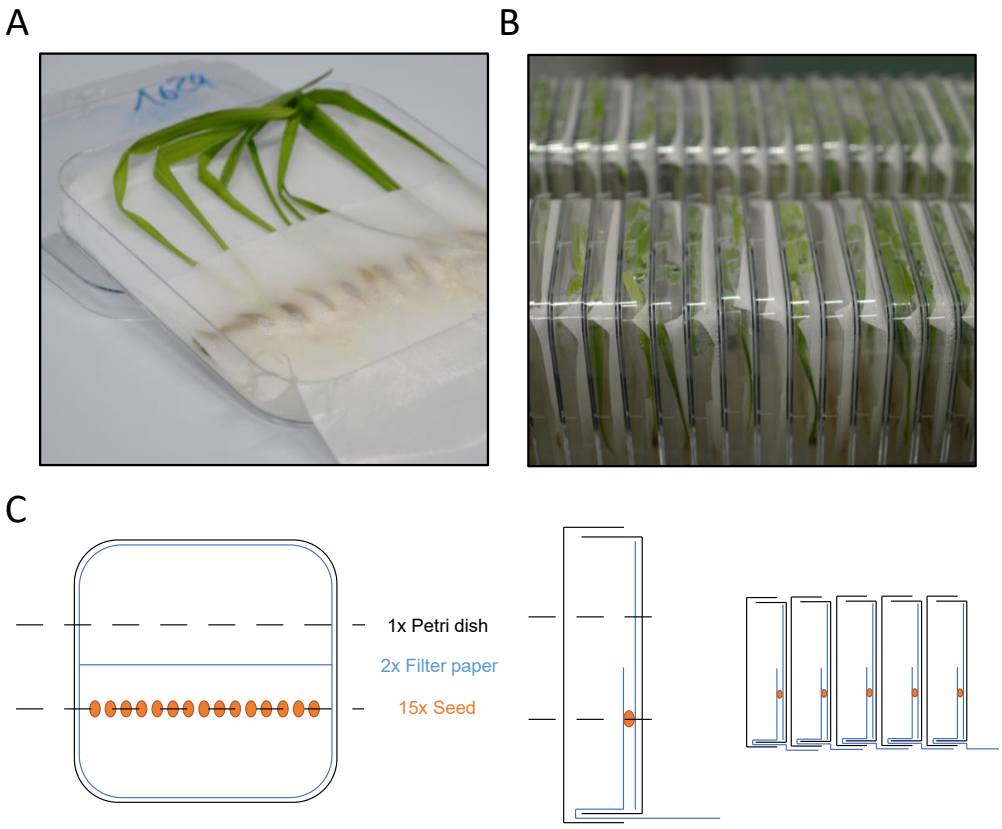


Fig. S1. Illustration of the cultivation system of (A) barley seedlings in Petri dishes using (B) space efficient cultivation. (C) Schematic overview of the approach used to cultivate barley seedlings in reach-in growth chambers.

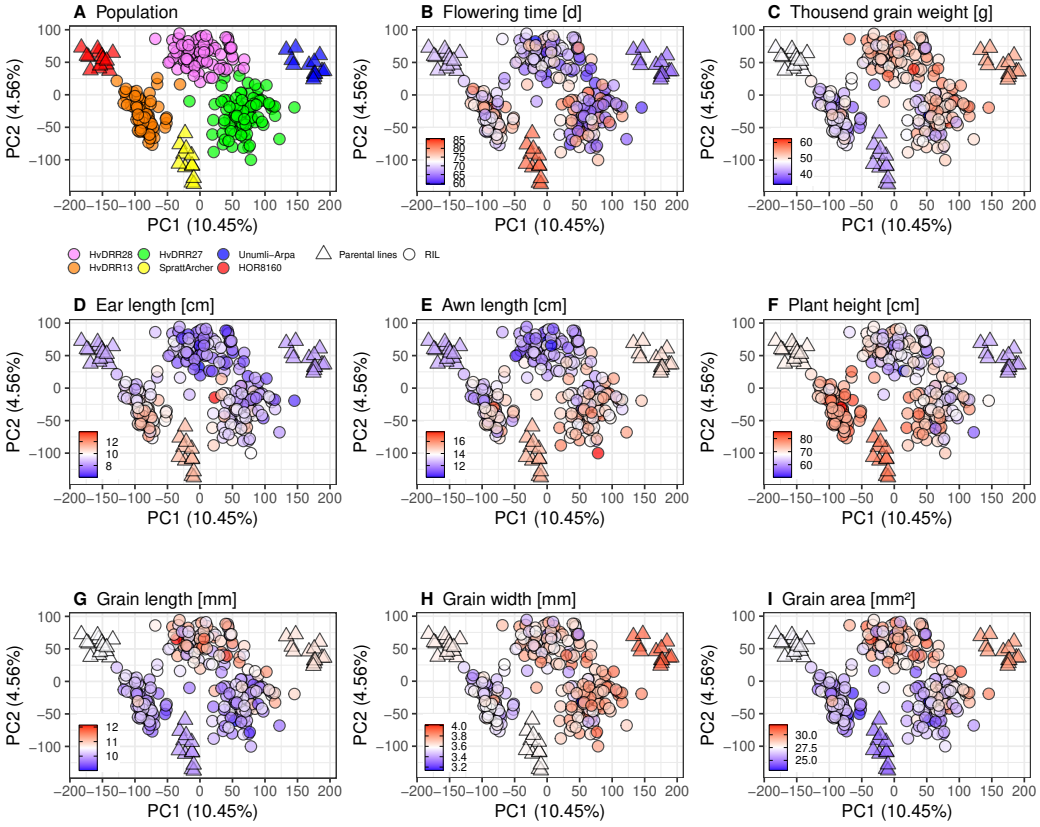


Fig. S2. Principal component analysis (PCA) based on GE^{DEG}_{RNAseq}, (A) overlayed with the information of the genetic material and (B-I) the adjusted entry means of all eight traits. Shape differentiate RIL and parental inbreds. PC 1 and PC 2 are the first and second principal component, respectively, and the number in parentheses refers to the proportion of variance explained by the principal components in percent.

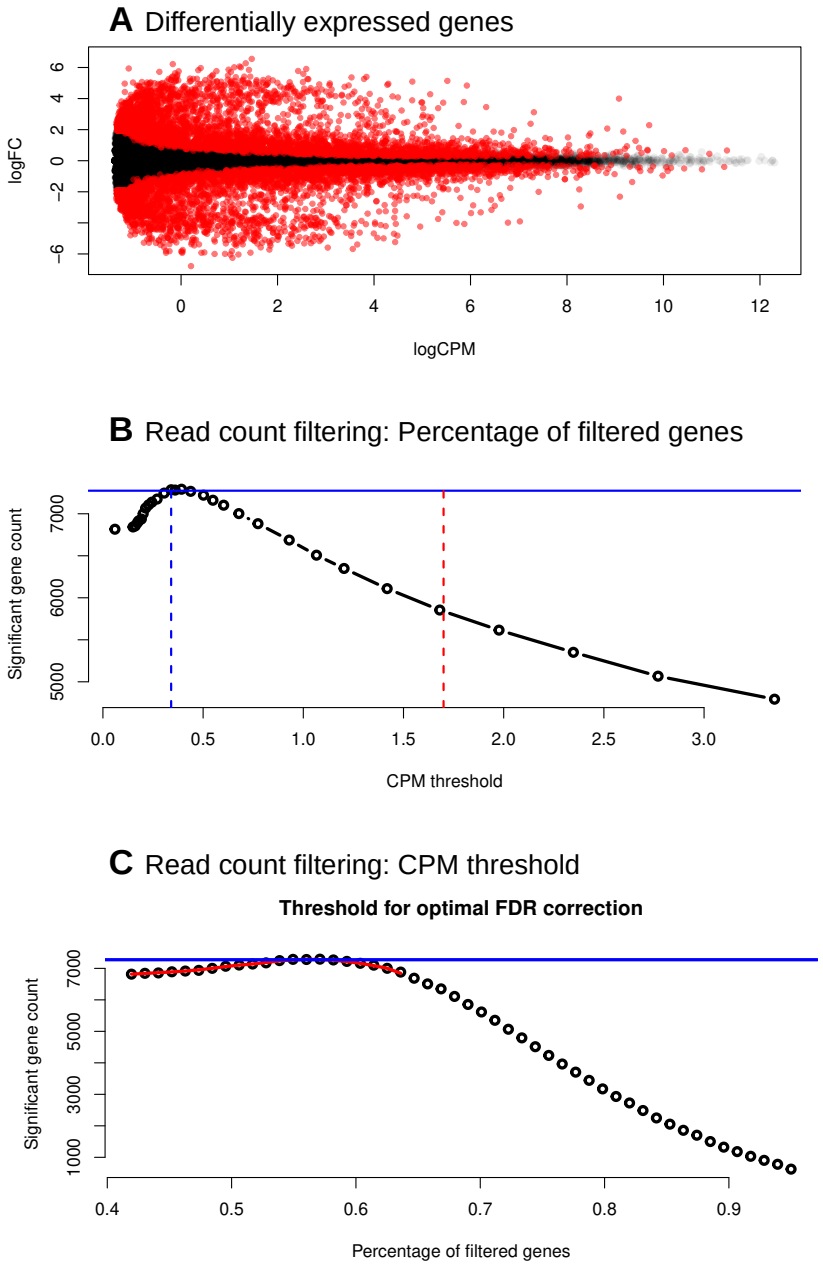


Fig. S3. overview of the read count filtering. (a) the number of differentially expressed genes shown in red could be maximized by adjusting the percentage of (b) filtered transcripts or (c) cpm threshold. the blue horizontal line indicating the maximum of differentially expressed genes counted. the dotted vertical lines show the chosen threshold (blue) and the edger standard threshold (red).

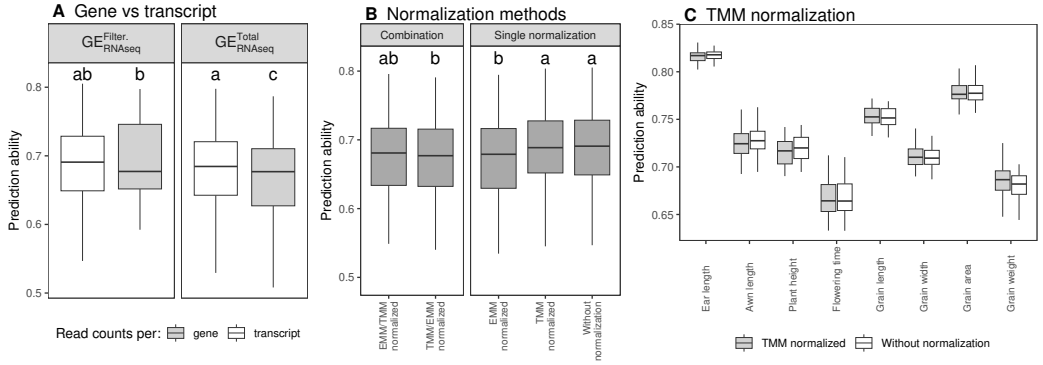


Fig. S4. Characterization of the expression data. (A) Comparison of the genomic prediction results for two different read counting methods: counting per gene and counting per transcript for GE^{Total}_{RNAseq} and $GE^{Filter.}_{RNAseq}$. (B) Genomic prediction results for GE^{DEG}_{RNAseq} compared based on the order of normalization: Trimmed Mean of the M-values (TMM) first or estimated marginal means (EMM) based on block effect first. (C) Genomic prediction abilities for GE^{DEG}_{RNAseq} comparing unnormalized and TMM normalized data for all eight traits.

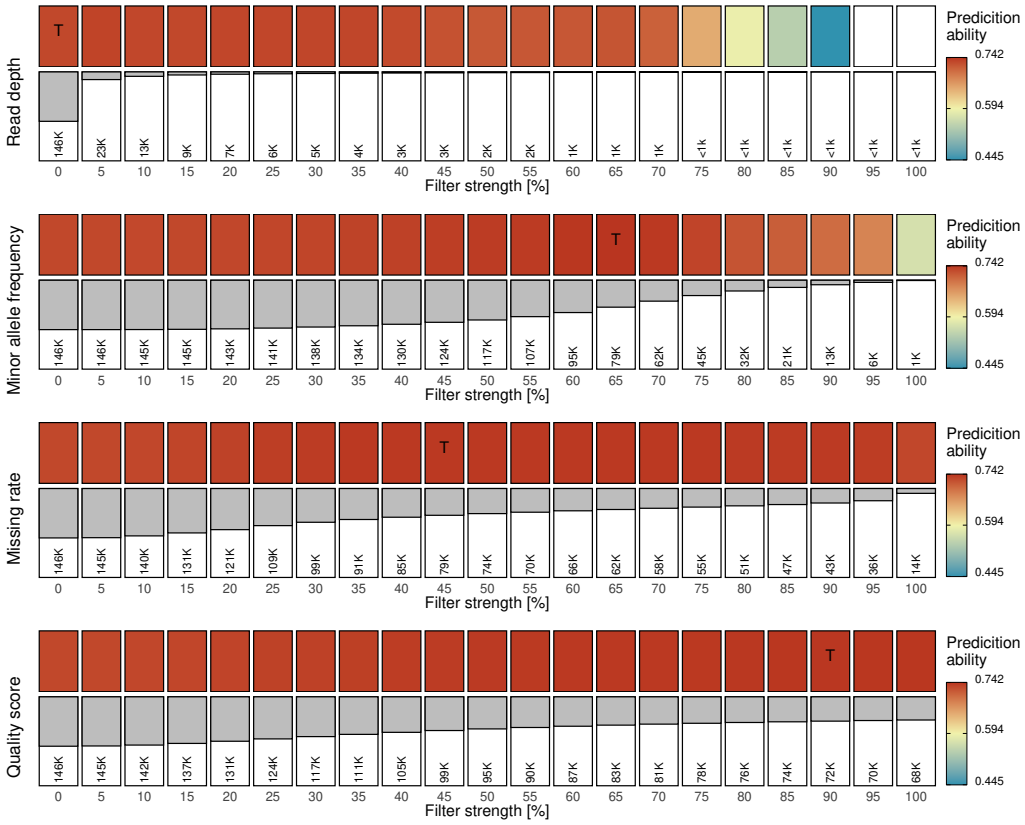


Fig. S5. Quality filtering of genomic variants of sequencing data based on genomic prediction (GP) performance. The performance shown are prediction abilities averaged across all eight traits. For each of the four criteria the prediction ability (top) and the remaining number of markers in thousands (bottom) are shown for 21 different relative filtering strength subsets from the minimum (0%) to the maximum (100%) value in the original dataset (SNP^{Total}_{RNAseq}). The best performing filter strength subset is marked (T). No GP was performed (white) when the number of remaining markers was insufficient.

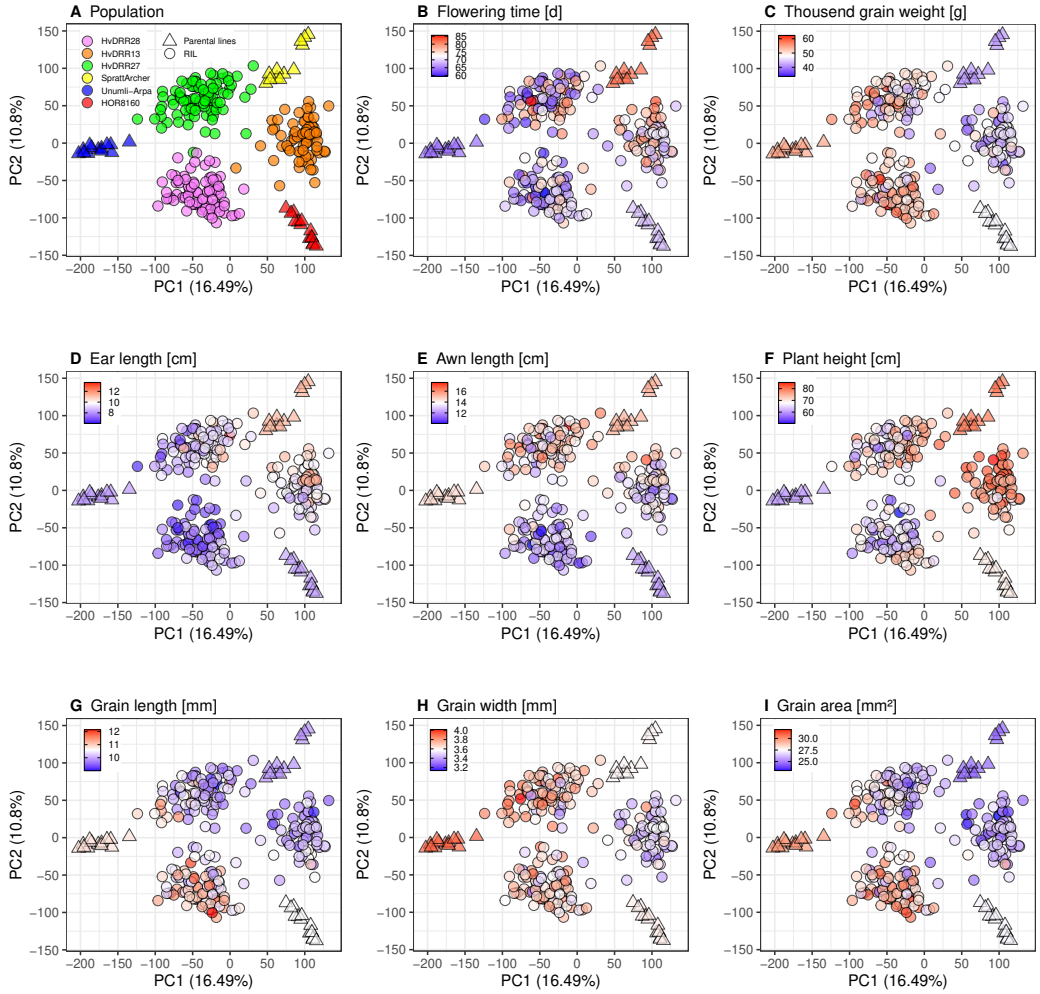


Fig. S6. Principal component analysis (PCA) based on SNP^{QC}_{RNAseq}, (A) overlayed with the information of the genetic material and (B-I) the adjusted entry means of all eight traits. Shape differentiate RIL and parental inbreds. PC 1 and PC 2 are the first and second principal component, respectively, and the number in parentheses refers to the proportion of variance explained by the principal components in percent.

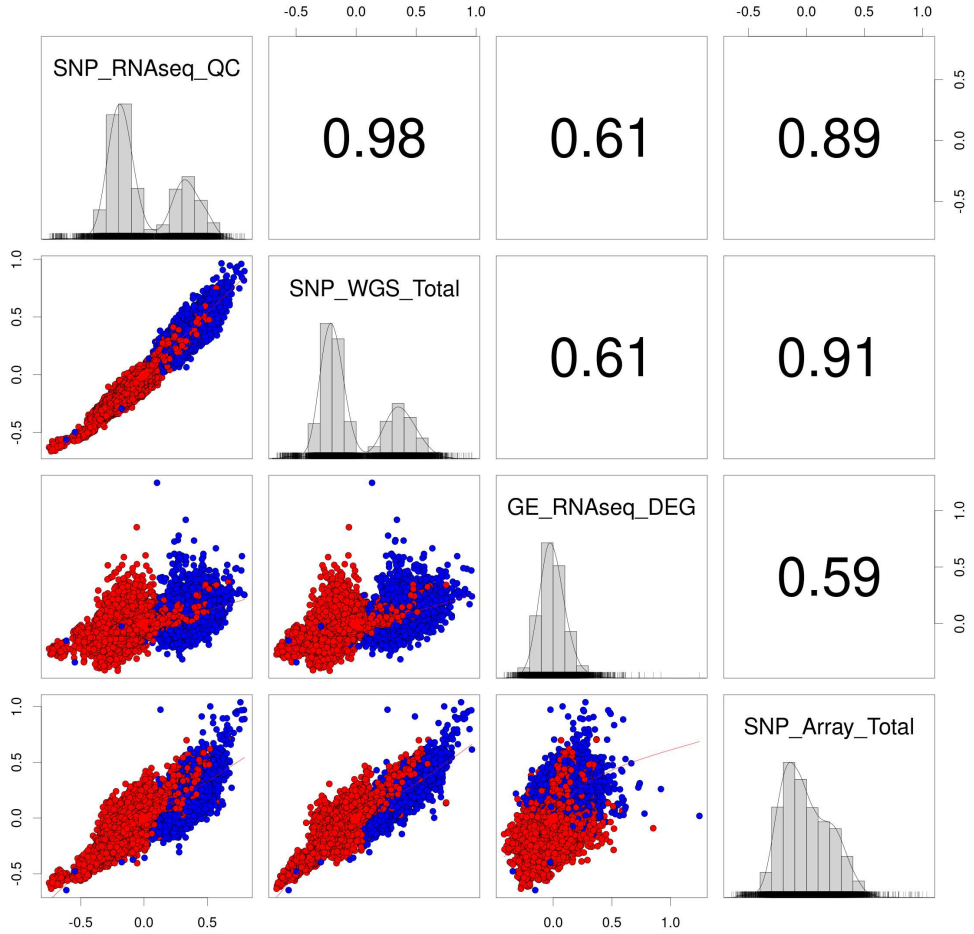


Fig. S7. Correlation plots of the additive relationship matrices of the datasets: SNP^{QC}_{RNAseq} (SNP_RNAseq_QC), SNP^{Total}_{WGS} (SNP_WGS_Total), GE^{DEG}_{RNAseq} (GE_RNAseq_DEG), SNP^{Total}_{Array} (SNP_Array_Total). The dots are colored by sample relationship, with intra-population comparisons blue and inter-population comparisons red.

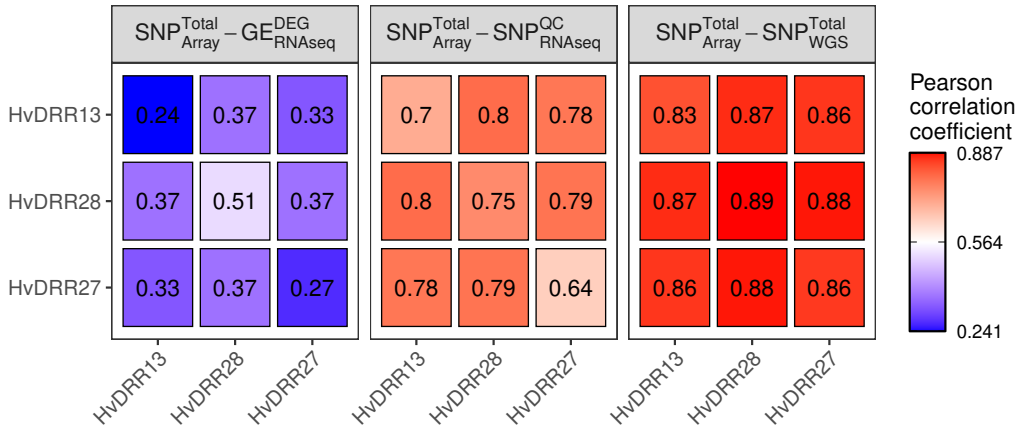


Fig. S8. Intra and Inter-population additive relationship matrices comparison. Heat map of Pearson correlation coefficients between the covariances of population segments of two additive relationship matrices. Comparison between SNP^{Total}_{Array} and the three remaining main datasets: SNP^{QC}_{RNAseq}, SNP^{Total}_{WGS}, and GE^{DEG}_{RNAseq}.

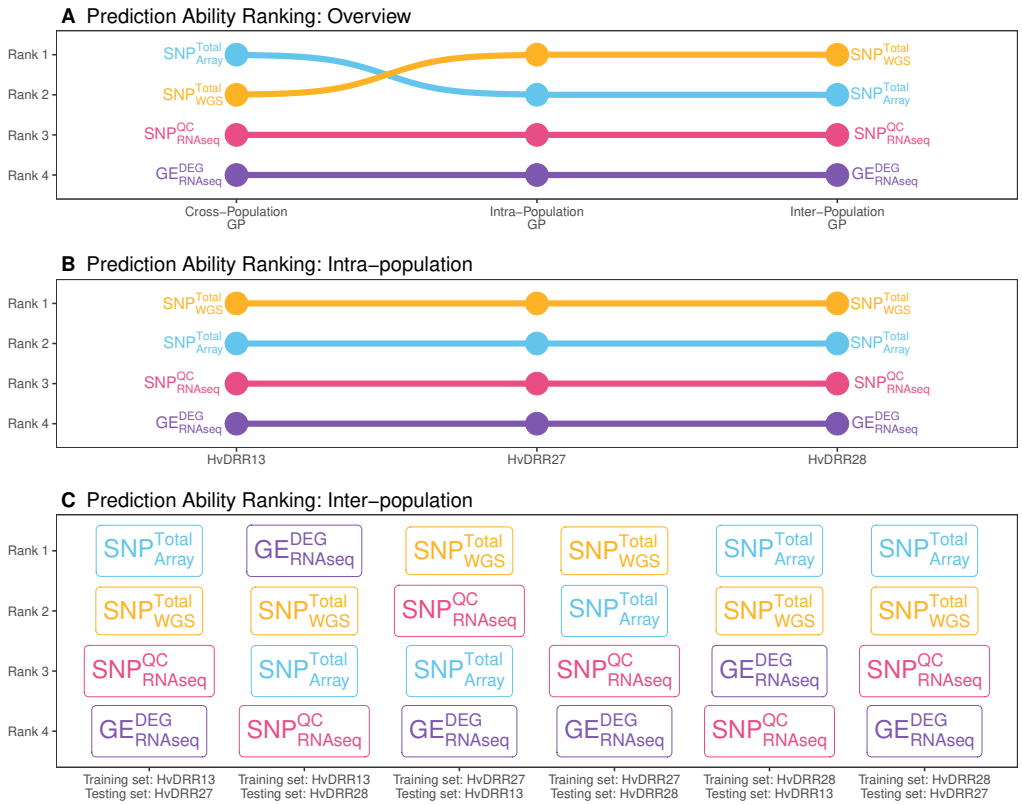


Fig. S9. Prediction Ability (PA) ranking for (A) intra-population and (B) inter-population comparisons using the best performing datasets from each data origin group.The training set size was set to 50 for all tests.

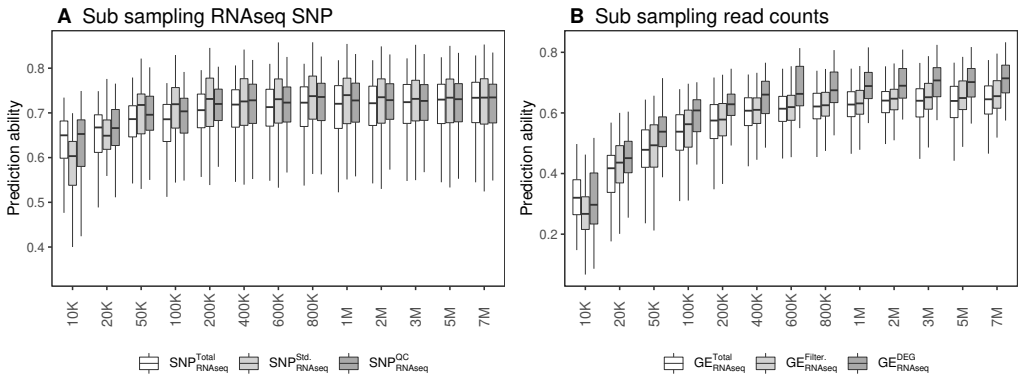


Fig. S10. Genomic prediction (GP) performance of artificially reduced sequencing depth subsets created by random read sub-sampling of (A) RNA-Seq genomic variant datasets (SNP^{Total}_{RNAseq}, SNP^{QC}_{RNAseq}) and (B) gene expression datasets (GE^{Total}_{RNAseq}, $GE^{Filter.}_{RNAseq}$, GE^{DEG}_{RNAseq}) including 155 of the 240 samples. The sub-sampling ranged in 13 steps between 10 thousand to 7 million reads.

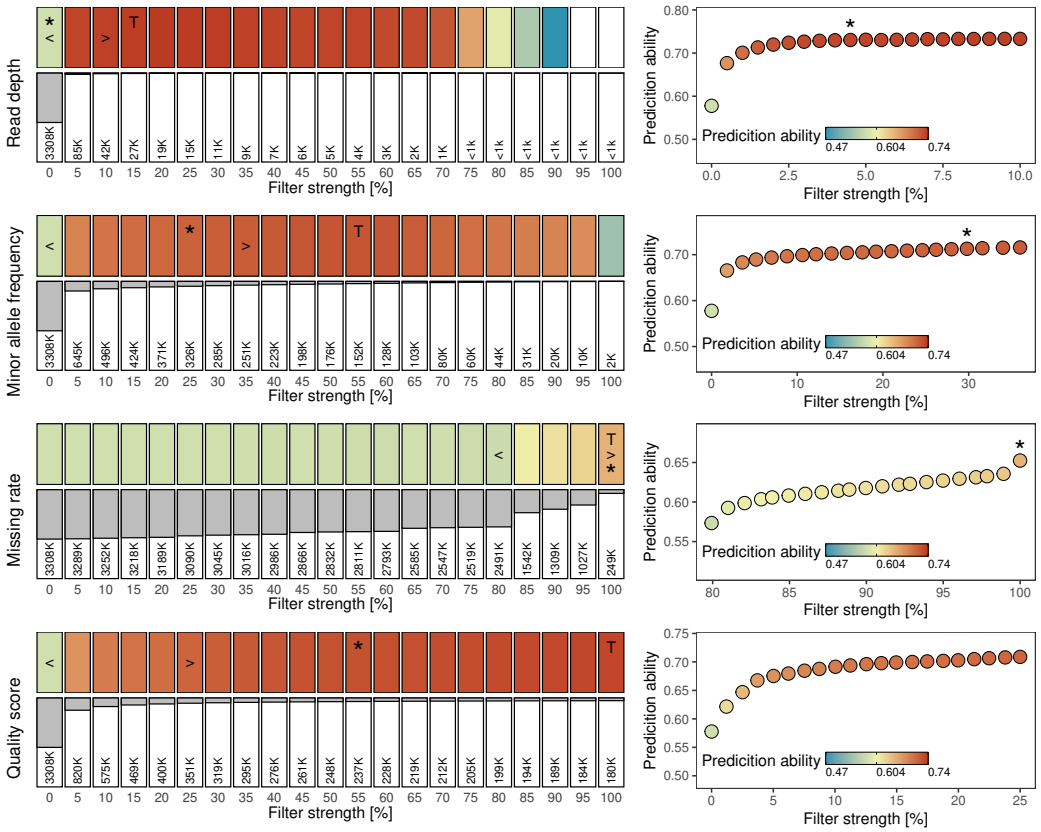


Fig. S11. Quality filtering of genomic variants of sequencing data based on genomic prediction (GP) performance. Quality filtering is an essential step in maximizing the GP potential and can have a larger effect on the results. We tested the impact of quality filtering on a less strictly cleaned dataset by creating SNP^{Raw}_{RNAseq}, an alternative version of SNP^{Total}_{RNAseq}, which included heterozygous / inconsistent allele calls and marker with missing parental data. From the 5.9M raw variants SNP^{Raw}_{RNAseq} included 3.4M, which is much more then SNP^{Total}_{RNAseq} (148K) and it resulted in a reduction GP performance by 0.15, but applying the same quality filtering workflow resulted in a subset with a comparable GP performance to SNP^{QC}_{RNAseq}.

The performance shown are prediction abilities averaged across all eight traits. For each of the four criteria the prediction ability (left, top) and the remaining marker in thousands (left, bottom) are shown for 21 different relative filtering strength subsets from the minimum (0%) to the maximum (100%) value in the original dataset (SNPTotal RNAseq). The best performing filter strength subset is marked top (T) as well as the optimal filter strength subset (*). The filtering subset marked * are not significantly ($p > 0.05$) different to T. The filter strength region with the highest impact on GP performance is shown in detail (right). The section is outlined on the overview as from start (<) to end (>).

6 The role of methylation and structural variants in shaping the recombination landscape of barley

Authors:
Federico Casale, **Christopher Arlt**, Marius Kühl, Jinquan Li, Julia Engelhorn, Thomas Hartwig, and Benjamin Stich.

Own contribution: Co-second author. I created and analyzed the RNA-Seq dataset and contributed to the manuscript in the corresponding sections.

The role of methylation and structural variants in shaping the recombination landscape of barley

Federico Casale[1], Christopher Arlt[1,2], Marius Kühl[1,2], Jinquan Li[3,§], Julia Engelhorn[3,4], Thomas Hartwig[3,4], and Benjamin Stich[1,2,3,5*]

[1]Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
[2]Julius Kuehn Institute, Federal Research Center for Cultivated Plants, Institute for Breeding Research on Agricultural Crops, 18190 Sanitz, Germany
[3]Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany
[4]Institute for Molecular Physiology, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
[5]Cluster of Excellence on Plant Sciences, From Complex Traits toward Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany
[*]Corresponding author: Benjamin Stich; benjamin.stich@julius-kuehn.de
[§]Current address: Strube D&S GmbH, Hauptstrasse 1, 38387 Söllingen

July 21, 2024

ABSTRACT

Meiotic recombination is not only a key mechanism for sexual adaptation in eukaryotes but crucial for the accumulation of beneficial alleles in breeding populations. The effective manipulation of recombination requires, however, a better understanding of the mechanisms regulating the rate and distribution of recombination events in genomes. Here, we identified the genomic features that best explain the recombination variation among a diverse set of segregating populations of barley at a resolution of 1 Mbp and investigated how methylation and structural variants determine recombination hotspots and coldspots at a high-resolution of 10 kb. Hotspots were found to be in proximity to genes and the genetic effects not assigned to methylation were found to be the most important factor explaining differences in recombination rates among populations along with the methylation and the parental sequence divergence. Interestingly, the inheritance of a highly-methylated genomic fragment from one parent only was enough to generate a coldspot, but both parents must be equally low methylated at a genomic segment to allow a hotspot. The parental sequence divergence was shown to have a sigmoidal correlation with recombination indicating an upper limit of mismatch among homologous chromosomes for CO formation. Structural variants (SVs) were shown to suppress COs, and their type and size were not found to influence that effect. Methylation and SVs act jointly determining the location of coldspots in barley and the weight of their relative effect depends on the genomic region. Our findings suggest that

recombination in barley is highly predictable, occurring mostly in multiple short sections located in the proximity to genes and being modulated by local levels of methylation and SV load.

Keywords: recombination rate, hotspot, methylation, structural variant.

4

INTRODUCTION

In sexual reproduction, the recombination between the maternal and paternal homologous chromosomes followed by their random assortment during meiosis generates new combinations of alleles that can be transmitted to the next generation (Barton and Charlesworth 1998). This mechanism for generating genetic diversity is widely conserved among eukaryotes since it provides a possibility for the adaptation of species: the reshuffling of alleles breaks the linkage between beneficial and deleterious mutations, allowing the accumulation of beneficial mutations into one haplotype, and, ultimately, generating new phenotypes upon which selection can act (Muller 1932; Peck 1994).

Meiosis consists of a single round of DNA replication in which the bivalent is generated –a pair of physically linked homologous chromosomes each composed of two replicated sister chromatids– followed by two rounds of chromosome segregation (for a review see Mercier et al. 2015). During the first round, meiotic recombination occurs in prophase I when programmed DNA double-strand breaks (DSBs) are repaired as either crossovers (COs), the reciprocal exchange of large regions between homologous nonsister chromatids, or noncrossovers (NCOs), –the unidirectional copies of small fragments from any of the intact homologous chromatids (Szostak et al. 1983). In addition, when both CO and NCO occur, mismatches at the site of the strand invasion are produced if sequence polymorphisms exist among homologous chromosomes. Such mismatches are either restored to the original

allelic state (i.e., intersister repair) or repaired in favor of the homologous allele resulting in gene conversion (GC): the nonreciprocal exchange of alleles between homologous nonsister chromatids (Burt 2000; Wijnker et al. 2013).

The rate of recombination events is strongly regulated and has been proposed to be related to an optimal level of recombination for adaptation in eukaryotes (Mercier et al. 2015). The presence of at least one CO per bivalent, termed the obligate CO, is required for the correct segregation of homologous chromosomes (Hall 1972). In addition, the number of generated DSBs exceeds the number of observed COs which are rarely more than three per chromosome per meiosis in most species (Martini et al. 2006; Baudat and Massy 2007; Bauer et al. 2013). Most of the observed COs, furthermore, prevail in small regions of a few kilobases called recombination hotspots where CO rates are several times greater than the chromosome average (Mézard 2006; Choi and Henderson 2015). Remarkably, both the rate and distribution of COs in the genome have been shown to exhibit extensive inter- and intraspecific variation (Nachman 2002; Ritz et al. 2017; Lawrence et al. 2017). The mechanisms behind such variation, however, are not completely understood. Such knowledge is required for the effective manipulation of recombination, e.g., for the purpose of plant breeding. This recombination determines the frequency of breaking undesirable linkages and stacking favorable alleles in the genetic background of breeding populations and defines marker resolution to map quantitative traits (Bauer et al. 2013; Blary and Jenczewski 2019).

Earlier studies provided valuable information for deciphering the mechanisms

behind recombination variation in plants. Most of the recombination in plants occurs in euchromatic regions where chromatin is accessible while heterochromatin is suppressed for meiotic recombination (Henderson 2012). For example, earlier studies in *Arabidopsis thaliana* and other species have shown that recombination events tend to occur in genomic regions with hypomethylated DNA (Yelina et al. 2012; Rodgers-Melnick et al. 2015; Marand et al. 2019; Apuli et al. 2020; Fernandes et al. 2024) and depleted nucleosome density (Choi et al. 2013; Wijnker et al. 2013). Moreover, COs in plants are typically located in close proximity to genes and in association with chromatin marks that favor transcription (Choi et al. 2013; Mercier et al. 2015). A positive correlation between the recombination rate and gene density has been observed in many plant families (Paape et al. 2012; Choi et al. 2013; Silva-Junior and Grattapaglia 2015; Gion et al. 2016; Wang et al. 2016; Apuli et al. 2020), including most grasses (Bauer et al. 2013; Darrier et al. 2017; Jordan et al. 2018; Gardiner et al. 2019; Marand et al. 2019; Casale et al. 2022). It is worth noting that some studies have instead reported a negative correlation between recombination rate and gene density (Kim et al. 2007; Giraut et al. 2011; Yang et al. 2012; Rodgers-Melnick et al. 2015).

Polymorphisms among homologous chromosomes are expected to prevent recombination due to defective strand invasion and homology pairing caused by the increase in mismatches among nonsister chromatids (Henderson 2012). Earlier studies in plants, however, reported constrasting correlations between recombination rate and parental sequence divergence (Saintenac et al. 2011; Salomé et al. 2012;

Yang et al. 2012; Bauer et al. 2013; Jordan et al. 2018; Marand et al. 2019). Recent reports in *Arabidopsis* suggested that the recombination rate has a positive correlation with parental allelic divergence until a level of mismatch among homologous chromosomes prevents CO formation (Blackwell et al. 2020; Hsu et al. 2022). Accordingly, large structural variants (SVs) were shown to suppress local recombination in several plant species (Rodgers-Melnick et al. 2015; Shen et al. 2019; Rowan et al. 2019; Fernandes et al. 2024).

The above-mentioned contrasting findings impose the necessity to characterize the associations between recombination and genomic features at the species level to avoid making incorrect assumptions. In addition, due to differences in the employed research methods, disagreements were observed among studies on the same species (Apuli et al. 2020). In this respect, most reported recombination rates in plants were calculated based on a coarse genomic resolution that failed to capture the complete genetic variation generated from meiosis. At present, most reported recombination assessments at high-resolution have been performed in *Arabidopsis* (Sun et al. 2012; Lu et al. 2012; Yang et al. 2012; Wijnker et al. 2013; Rowan et al. 2019; Fernandes et al. 2024), while only a few have been performed in major cereal crops such as maize (*Zea mays*) (Rodgers-Melnick et al. 2015; Li et al. 2015), rice (*Oryza sativa*) (Si et al. 2015; Marand et al. 2019), and wheat (*Triticum aestivum*) (Jordan et al. 2018). In barley (*Hordeum vulgare*), the fourth of this list, earlier low-resolution recombination studies successfully revealed the genetic basis of recombination as well as the association of recombination with some environmental and genomic

features on a broad genomic scale (Higgins et al. 2012; Dreissig et al. 2019, 2020; Casale et al. 2022); however, a high-resolution study depicting the complete meiotic recombination landscape and the respective associations of genetic and epigenetic features in the barley genome is still lacking.

Consequently, in the present study, we aimed to (*i*) identify the genomic features that best explain the recombination variation among the double round-robin (DRR) populations, (*ii*) detect recombination events in the barley genome at high-resolution, and (*iii*) analyze the effect of genomic features in determining the location of recombination hotspots and coldspots in the genome.

RESULTS

The genomic features associated with recombination rate variation in the barley genome

The recombination rate was almost null in the pericentromeric region, but increased toward the distal regions of the chromosomes (Figure 1A). The SV load, the physical fraction spanned by the analyzed SVs (insertions, deletions, inversions, duplications, and translocations) increased toward the distal regions of the chromosomes as did the recombination rate and gene density (Figures 1A and S1). In this way, the presence of SVs and the sequence divergence among parents showed a significant positive Pearson's correlation ($P < 0.05$) of 0.35 with the recombination rate along the distal regions of the chromosomes (Figures 1C and S3A), whereas a lower but significant correlation ($P < 0.05$) of 0.1 was observed in the pericentromeric region (Figures 1D and S3B). The sequence divergence among the parental inbred lines, which showed a significant positive correlation ($P < 0.05$) with the SV load of 0.27, was found to have a significant positive correlation ($P < 0.05$) with recombination rates of only 0.1 and 0.14 in the distal and pericentromeric regions, respectively (Figures 1A, 1C, and 1D). The methylation level in the sequence contexts differed along the barley chromosomes, where the level in the CpG and CHG contexts reached a maximum in the pericentromeric region and decreased toward the telomeres, while the methylation level in the CHH context increased toward the telomeres (Figure 1B). Because the CpG and CHG contexts had higher overall

methylation levels than did the CHH context, the average methylation level along the chromosomes mostly represented the trend observed for the CpG and CHG contexts. The observed significant negative correlation ($P < 0.05$) between the average methylation level and recombination rate along the barley chromosomes was therefore due to the CpG and CHG contexts but not to the CHH context (Figures 1C and 1D). The difference in the methylation level between the parental inbred lines of any of the analyzed populations was greater in the distal regions than in the pericentromeric regions of the chromosomes in all three analyzed sequence contexts (Figure S4). This explained the positive correlation ($P < 0.05$) of 0.17 of such difference and the recombination rate along the distal regions in the barley chromosomes (Figures 1C, 1D, S3A, and S3B).

The genomic features associated with recombination rate variation among barley populations

The best subset of genomic features explaining the recombination rate variation among the 45 DRR populations in 1 Mbp genomic windows along the barley chromosomes was identified using a stepwise regression model (Figure 2). The fraction of 1 Mbp windows of the barley genome in which a given genomic feature was found to be significantly correlated and the direction of such correlation are given in parentheses below. The genetic effects, which were calculated from the GRE, were found to be the most determining factor in explaining differences in recombination rate among populations with a promoting effect on both the distal (positive, 0.76) and

the pericentromeric regions (positive, 0.40) of the chromosomes. The parental sequence divergence was also found to be positively correlated with the recombination rate in both mentioned chromosomal regions (positive, 0.24 and 0.19, respectively). A few windows showed a significant negative correlation between parental sequence divergence and the recombination rate (negative, 0.04 and 0.01 in both respective chromosomal regions). Notably, the windows with a negative correlation had a parental sequence divergence near its relative maximum, which appeared to be associated with a high SV load (Figure S5A). The average methylation level across the different sequence contexts was negatively correlated with the recombination rate in the distal (negative, 0.43) and pericentromeric (negative, 0.12) regions of the barley chromosomes. This means that populations with a higher methylation level than others in a particular window showed a lower recombination rate than others in that window, and vice-versa. Additonally, the multiple regression model used to calculate genetic effects revealed a correlation between the methylation level and the sum of the GREs ranging from -0.29 to 0.03 per 1 Mbp genomic window with an average of -0.06 in the distal region. The difference in the average methylation level between the parental inbred lines of the respective populations was found to have only a low impact on the recombination rate variation among populations (in distal regions, positive, 0.04; negative, 0.03). The physical fraction of 1 Mbp genomic windows spanned by all SVs was found to have a low (and mostly positive) impact on the differences in recombination rate among populations (in distal regions, positive, 0.05; negative, 0.02).

The methylation level in the three sequence contexts –CpG, CHG, and CHH– analyzed independently in an extended model was shown to have the same repressive effect on recombination as the average across the three sequence contexts (Figure S6 and Table S2). In addition, the parental methylation difference for the sequence contexts CHG and CHH was found to be positively associated with the recombination rate, and the inverse was found for the context CpG. Furthermore, the extended model revealed no impact of the different SV types on recombination rate variation among populations.

The key parameters for detecting recombination events at high-resolution

The mRNA sequencing of recombinant inbred lines (RILs) of the evaluated populations yielded 4.1 M sequence variants of which, 858 K SNPs remained after being intersected with reported SNV parental data generated by DNA sequencing. A total of 12 K SNPs were added from the iselect array, resulting in a total of approximately 870 K SNPs genome-wide. After reoving SNPs carrying nonparental alleles or with 100% missing data, the final number of genome-wide SNPs ranged from 214 to 259 K per population (Table S3 and Figure S7). The median inter-SNP distance was 132 bp on average across populations, which was 79 times shorter than the $10,475$ bp utilized to count COs for the same three populations analyzed in a previous study (Casale et al. 2022). The maximum inter-SNP distance ranged from 5.45 to 11.99 Mbp among populations (Table S3), denoting large regions that

were identical by descent (IBD) among the parental inbred lines involved in a given population and thus among their respective offspring. This resulted in an average density of 57 SNPs per Mbp across populations. The mean length for each parental SNP block category was 14.8, 44.1, and 111.5 Mbp for the short, medium, and long CO-related blocks, respectively (Table S4). The block length was positively correlated with the number of SNPs per block (> 0.75, $P < 0.001$). Therefore, the false positive rate for detecting CO was inversely proportional to the marker block length, supporting the identification of block length categories with different CO layers.

On average, there were were 30, 87, and 269 genome-wide COs accumulated per RIL across populations for the long, medium, and short block lengths, respectively (Table S5). Considering only the layer of COs generated by blocks longer than 3 Mbp, the genome-wide CO counts per RIL ranged from 14 to 65 across populations (Figures S8 and S9). Since a given CO breakpoint was determined as the midpoint of the CO interval, the breakpoint location accuracy depended on the CO interval length (Table S6). The lengths of the detected CO intervals ranged from less than 20 bp to 10.8 Mbp with a median that varied from 41.9 kb to 151.6 kb depending on the considered CO layer. The average number of genome-wide GC events that accumulated per RIL across populations was 58, 251, and 6,521 for long, medium, and short GC-related block lengths, respectively.

An average of 80 recombination hotspot windows per chromosome were found across the three selected populations (Table S7). Among these, 12 were found

in the pericentromeric regions, and the rest were found in the distal regions of the barley chromosomes (Table S8 and Figure S10). The recombination hotspot windows contained 0.26, 0.14, and 0.25 of the total observed COs in the HvDRR13, HvDRR27, and HvDRR28 populations, respectively. Less than 10% of the hotspot windows in a given population were shared with another population and less than 1% of the total counted hotspot windows were shared among the three analyzed populations (Figure S11). Interestingly, both the number and conservation level of coldspot windows far exceeded those of hotspots. On average, across populations, more than half (10,436 out of the 19,496) of the distal windows were recombination coldspots. More than 60% of the coldspot windows in a given population were shared with the other populations, and 16.7% of the coldspot windows were present in all three analyzed populations. The majority of the coldspot windows were located contiguously in regions with lengths that varied from 10 kb to 17 Mbp with an average of 322 kb across the three analyzed populations.

More than 15% of the GC hotspot windows in a given population were shared among the three analyzed populations (Figure S12). The GC hotspot windows were found to overlap with the CO hotspot windows in the distal region of the barley chromosome significantly more than they did under a random distribution across the three analyzed populations (Table S9). The GC hotspot windows detected in a given population overlapped with 12–15% of the CO hotspot windows in the same population and with 7.5–9.5% of the CO hotspot windows detected in the other two analyzed populations (Table S10). There was no significant ($P > 0.05$) difference

between such overlap proportions in the HvDRR27 and HvDRR28 populations.

The effect of methylation and structural variants on recombination rate variation at high-resolution

The coldspot and hotspot windows have different methylation level and SV load than the rest of the genome

The coldspot and hotspot windows in a given chromosome region showed distinct methylation patterns compared to the average remaining windows in the same region in the three analyzed populations (Figure 3). The average methylation level across all three methylated sequence contexts in the coldspot windows of the distal telomeric subregion was significantly greater ($P < 0.001$) than that across the other windows in both distal subregions. In contrast, the coldspot windows in the distal proximal region were not found to be differentially methylated ($P > 0.001$) from other windows in any of the distal subregions. However, when analyzing the CpG and CHG sequence contexts separately, the methylation level in the coldspot windows of the distal proximal subregion was found to be significantly greater ($P < 0.001$) than that in the other windows of both distal subregions. Differently, the methylation level in the coldspot windows of the distal telomeric subregion was significantly greter ($P < 0.001$) than that in the other windows in this subregion but significantly lower ($P < 0.001$) than that in the windows of the distal proximal subregion. The coldspot windows in such comparisons that were below the critical value (methylation levels of 0.89 and 0.59 for the sequence contexts CpG and CHG,

respectively) were found to have a significantly ($P < 0.001$) greater total SV load fraction than the coldspot windows above the critical value (Table S11).

The average methylation level across the three sequence contexts in the hotspot windows was significantly lower ($P < 0.001$) than that across the other windows in any region of the barley chromosomes. The hotspot windows in the pericentromeric region, however, were not found to be differentially methylated from the rest of the windows in such region or from the windows in other regions of the genome, including coldspots. However, by analyzing the methylated sequence contexts separately, the hotspot windows were found to be significantly less methylated ($P < 0.001$) in the CpG and CHG sequence contexts than in the total windows in the pericentromeric region.

The coldspot windows in the distal telomeric regions were found to have a significantly greater total SV load ($P < 0.001$) than the rest of the windows in that subregion. In contrast, the coldspot windows in the distal proximal region did not show such an increase in total SV load. However, the coldspot windows below the critical value (SV loads equal to 0.187, 0.174, and 0.187 for the HvDRR13, HvDRR27, and HvDRR28 populations, respectively) in such comparisons had a significantly increased methylation level ($P < 0.05$) compared to the windows above the critical value for the CpG and CHG sequence contexts (Table S12). The hotspot windows were not observed to have a significantly different ($P > 0.001$) span of total SVs compared to the total windows in their respective chromosome regions. However, the observed overlaps between CO intervals and insertions/deletions and

duplications were found to be significantly less frequent ($P < 0.001$) than such overlaps under a random distribution of the COs and the respective SVs in the distal regions of the barley chromosomes in the three analyzed populations (Table S13). In the case of inversions, such a pattern was observed only for the HvDRR27 population. Moreover, the distance between the CO breakpoints and the closest SV of any type was significantly greater ($P < 0.001$) then the CO-SV distances expected by chance (Table S14).

The genomic environment neighboring hotspot and coldspot windows

The 10 kb genomic windows adjacent to coldspot regions were found to have a significantly lower ($P < 0.001$) average methylation level across the three sequence contexts than coldspots in both distal subregions of the chromosome in the three analyzed populations (Figure 4). This observation reflected the pattern produced at the methylated sequence contexts CpG and CHG, individually (Figure S13). In addition, any 10 kb window in the considered range from -40 kb to +40 kb around coldspot regions was found to have a significantly lower ($P < 0.001$) total SV load than the neighboring coldspot. The 10 kb genomic windows adjacent to hotspot regions were not found to have significantly ($P > 0.001$) different methylation levels or SV loads than any of the analyzed chromosomal regions or populations.

The windows neighboring coldspot regions were found to have a significantly lower ($P < 0.001$) gene density than these regions, except for the windows located 20 kb upstream of coldspots. However, the overlap between the coldspot regions

and genes in the distal regions of the barley chromosomes was not significantly different ($P > 0.001$) from such overlap under a random distribution (Table 1). In contrast, a visual increase in the gene density from the hotspots to 20 kb upstream was observed in all the analyzed genomic regions, although this pattern was not significant ($P > 0.001$). Furthermore, the overlap between the hotspot regions and genes was found to be significantly ($P > 0.001$) greater than expected under a random distribution in the three analyzed populations, while the overlap between hotspots and intergenic regions was not found to be significantly ($P < 0.001$) greater than random, with the exception of the HvDRR27 population. In addition, a high proportion of the windows surrounding hotspot regions in both the proximal (0.37–0.49) and telomeric (0.33–0.45) subregions of the distal region of the genome were coldspot windows (Table S15).

The variation in methylation and SVs in coldspot and hotspot windows among barley populations

Significant differences ($P < 0.016$) were observed in the methylation levels of the three analyzed populations, either by analyzing the methylated sequence contexts separately or by analyzing their average (Figure 3). Such differences among populations observed for the total windows were also detected in the coldspot windows in all of the analyzed chromosome regions. In contrast, the methylation level in the hotspot windows was found to be equal ($P > 0.016$) among populations in any of the analyzed chromosome regions, either by analyzing the methylated sequence

contexts separately or their average. A similar trend was observed for the total SV load fraction: while observing significant differences ($P < 0.016$) among populations in the total windows but also in the coldspots of both the pericentromeric region and the distal subregions, such differences were not observed for the hotspot windows.

The methylation level of the two parental inbred lines of each of the analyzed populations differed significantly ($P < 0.008$) at the CpG and CHG sequence contexts of the genomic windows identified as coldspots in their respective offspring (Table 2). Thus, the increased methylation of only one of the parental genotypes at a genomic region might be enough to generate a coldspot in the offspring. In hotspot windows, no significant differences ($P > 0.008$) between parental inbred lines were found in the methylation level at any of the three analyzed sequence contexts, indicating that parents must have equally low methylation at a genomic segment to allow a recombination hotspot.

DISCUSSION

Detection of recombination events at high-resolution in barley

The substantial decrease of the median inter-SNP distance compared to a previous study with the same three populations (Casale et al., 2022) produced a slight increase of the CO discovery rate of 0.31 times when considering the COs related to $>$ 3 Mbp marker blocks. However, such increase jumps to 2.74 and 10.59 times, if considering the COs related to $>$ 500 kb and $>$ 10 kb blocks, respectively (Tables S3 and S5). The assumed part of the observed increase due to the additional recombination that occurred at heterozygous regions in the selfing generation analyzed in the first study is expected to be small due to the decreasing remaining heterozygosity after every selfing generation that produces fewer new observable COs per generation during inbreeding (Esch et al. 2007). By analyzing the detected CO rate between comparable low and high-resolution analyses reported in previous studies, only small differences were detected in populations of *Arabidopsis thaliana* and maize (*Zea mayz*) (McMullen et al. 2009; Rodgers-Melnick et al. 2015), but substantial differences were reported in populations of wheat (Esch et al. 2007; Gutierrez-Gonzalez et al. 2019; Gardiner et al. 2019) and *Populus* (Apuli et al. 2020). In addition to the different utilized resolutions, other reasons behind observing differences in the recombination rate in the same population may include the genotyping error rate, the data filtering criteria, and the size of considered CO events.

The observed CO rate per RIL per chromosome per generation in the present study, when considering the COs related to > 3 Mbp marker blocks, is in line with high-resolution studies in *Arabidopsis* (Sun et al. 2012; Lu et al. 2012; Yang et al. 2012; Wijnker et al. 2013; Qi et al. 2014; Rowan et al. 2019) and rice (*Oryza sativa*)(Si et al. 2015) that reported rates of 1.5–2.2 COs per chromosome per generation, and with another high-resolution study on wheat (Gardiner et al. 2019) when looking at the COs related to > 500 kb marker blocks.

The COs per RIL per chromosome per generation observed in our study when considering all COs related to blocks (10 kb) should be compared with values observed by Yang et al. (2012) in *Arabidopsis* and Gardiner et al. (2019) in wheat. In such comparisons, however, considering every marker block shorter than 10 kb as a GC is an arbitrary threshold. This has the potential to cause misclassification between COs and GCs among the categories of COs related to blocks > 10 kb and those related to GCs between 2 and 10 kb.

The present study is the first to characterize GC events in barley, along with a few in other crop species (Li et al. 2015; Si et al. 2015; Gardiner et al. 2019). The poor documentation of GCs in plants beyond studies in *Arabidopsis* is because phenotypic screens can barely detect GC events. In addition, the detection of GC events at the molecular genetic level is also challenging because of their short length (Mancera et al. 2009; Mercier et al. 2015). This makes GC detection very sensitive to the marker density and GC rate, which also depend on the recombination rate, the tract length of the repair intermediates where GCs occur, and the polymorphism

density (Wijnker et al. 2013). Moreover, in most flowering plants, gametes do not remain grouped after meiosis, making it difficult to observe the expected 3:1 allelic proportion of GCs (Sun et al. 2012). In the present study, the average number of genome-wide GC events per RIL across populations was 58, 251, and 6,521 for long (2–10 kb), medium (20 bp–2 kb), and short (2–20 bp) GC-related marker block lengths, respectively. The SNPs analyzed per population could be translated to 0.00003, 0.00014, and 0.0039 GC events per site per RIL per generation for the different types of GC-related marker block lengths. Marker blocks shorter than 20 bp are expected to contain a high number of false-positive GCs since they were predominantly called based on two markers only. Therefore, if considering the GC-related marker blocks of long (2–10 kb) and medium (20 bp–2 kb) length only, the detected 0.00017 GCs per site per RIL per generation is on the same order of magnitude as that observed in studies using a similar approach for GC detection (Yang et al. 2012; Gardiner et al. 2019). However, the observed GC rate in our study was two orders of magnitude greater than that reported in tetrad analysis studies performed in *Arabidopsis* (Lu et al. 2012; Sun et al. 2012; Wijnker et al. 2013) and sequencing of rice F2 populations (Si et al. 2015). This disagreement might be explained not only by the less precise GC detection method used in our study but also by the occurrence of nonallelic sequence alignments caused by SVs inflating the number of false-positive gene conversions (Qi et al. 2014; Si et al. 2015). It is also worth noting that the reported GC rate is the frequency of GCs generated from NCO and CO events combined. In the present study, we did not attempt

to estimate the rate of NCO in barley because these NCOs are only traceable after gamete formation when they lead to GC, and it was not possible to precisely differentiate between CO and NCO conversion tracts with the employed marker resolution. In addition, the DSB rate in barley is not known enough to estimate the NCO rate from the detected CO events.

The present study is the first to report genome-wide recombination hotspots at high-resolution in barley. On average, across the three investigated populations, of the 80 hotspots per chromosome, only 12 were found in the pericentromeric region, and the rest were found in the distal regions of the barley chromosomes (Figure S10 and Table S8). In addition, while the three investigated populations always shared one parent, the proportion of shared hotspots between two and three populations was 10% and 1%, respectively, of the total hotspots detected in a given population (Figure S11A). This observation is in good agreement with previous studies in *Arabidopsis* showing that recombination hotspots were cross-specific (Salomé et al. 2012).

In contrast, in the case of the GC hotspot windows, the overlap among the three populations was more than 15% (Figure S12). Moreover, GC hotspots were found to have high overlap not only with GC hotspots of the same population but also with windows that are hotspots in other populations (Table S10). Thus, GC hotspots might be considered fingerprints of population-specific silenced COs that result in NCO in regions with high DSB rates in the genome of the species. This observation suggested that although the CO rate and distribution present extensive intraspecies

variation, such DSBs might be highly conserved within the species. However, this requires further research.

Additionally, in barley, recombination hotspots alternate in the genome with coldspots. For example, in domains where CO rates are significantly lower than the genome average (Figure S10), as observed in previous studies in other species (Mercier et al. 2015). Indeed, by dividing the genome into 10 kb genomic windows, hotspot windows were found to be adjacent to coldspot windows in 42.5% of the cases in the distal regions of the barley chromosomes (Table S15). This continuos intermittence in the recombination rate might explain the above-mentioned large difference in CO events found between the high- and low-resolution analyses on these barley populations.

To avoid calling recombination coldspots in the pericentromeric and telomeric regions of the chromosomes, which are long regions depleted from recombination as seen in previous studies (Boideau et al. 2022), in the present study, coldspots were identified only in the distal region of the chromosomes by employing a long physical distance margin between regions. The majority of the detected 10 kb coldspot windows were located in coldspot regions with an average length of 322 kb (Table S7). Each population shared 60% of its coldspot windows and more than 16% with the other two populations, thus demonstrating a greater conservation of coldspots than hotspots in barley (Figure S11B). Such differential conservation between recombination hotspots and coldspots might be linked to the different genomic features determining their occurrence.

The genomic features associated with recombination rate variation in barley

The present study is the first comprehensive evaluation of the genomic features associated with differences in recombination rates among barley populations. On a scale of 1 Mbp windows, the recombination rate was found to be positively correlated with sequence divergence among parental inbred lines, gene density, and SV load on the barley chromosomes and was negatively correlated with the methylation level (Figure 1).

The results of the present study are in line with earlier studies in plants in which recombination was found to be positively associated with genetic divergence among homologous chromosomes (Yang et al. 2012; Marand et al. 2019; Blackwell et al. 2020). Although a negative association was reported in some studies (Saintenac et al. 2011; Gion et al. 2016; Bouchet et al. 2017; Serra et al. 2018; Jordan et al. 2018; Gutierrez-Gonzalez et al. 2019), the contradiction might be explained by a sigmoid relationship between both variables, meaning that recombinatiom has a positive correlation with genetic divergence until a level after which the high polymorphism among homologs suppresses COs due to the increase in mismatches, as recently reported in *Arabidopsis* (Blackwell et al. 2020; Hsu et al. 2022). In this respect, the few observed 1 Mbp windows with a significant negative association were found to have parental sequence divergence at the relative maximum level, which appeared to be associated with an extensive SV load (Figure S5A and B).

The SV load was not identified as a determining factor for the differences in the recombination rate among populations, presumably because the employed resolution of 1 Mbp was too coarse to detect differences among populations, as most of the analyzed SVs were smaller (Figure 2). Additionally, the positive correlation between the recombination rate and the SV load on a broad scale might be explained by the accumulation of DNA repair errors in highly recombining regions throughout the evolutionary history of barley (Figures 1A and 1C). Genomic regions with a high rate of DSBs are expected to have a historically increased mutation rate produced by COs and GCs, which elevates the allelic diversity at such regions among genotypes as demonstrated in humans (Arbeithuber et al. 2015; Halldorsson et al. 2019).

In addition to the positive correlation shown between recombination and gene density on a broad scale, in the present study, 10 kb hotspot windows were found to be associated with regions of high gene density (Table 1), as repeatedly reported in previous studies in grasses (Rodgers-Melnick et al. 2015; Darrier et al. 2017; Bouchet et al. 2017; Jordan et al. 2018; Gardiner et al. 2019; Marand et al. 2019; Casale et al. 2022), and other plant families (Paape et al. 2012; Choi et al. 2013; Silva-Junior and Grattapaglia 2015; Gion et al. 2016; Wang et al. 2016; Apuli et al. 2020). Furthermore, the hotspot windows were located in proximity ($<$ 20 kb apart) but did not overlap with the genes (Figure 4). This finding in barley is in line with previous observations in *Arabidopsis*, maize, and rice showing an increased CO frequency toward gene promoters and terminators (Choi et al. 2013; Wijnker

et al. 2013; Li et al. 2015; Marand et al. 2019; Sun et al. 2019), similar to that observed in budding yeast (*Saccharomyces cerevisiae*) (Pan et al. 2011).

In the present study, the genetic effects were calculated as the proportion of the sum of the *GRE* of both parents for a given population that was not explained by methylation, assuming parental sequence divergence and SV load as part of the *SRE* (Casale et al. 2022). Such genetic effects were shown to be the factor explaining the greatest proportion of differences in recombination rates among barley populations. Here, we hypothesize that such genetic effects are the product of the expression of genes related to the recombination machinery being in part modulated by the methylation level, explaining a portion of the uneven distribution of the recombination hotspots along the barley chromosomes. The wide distribution of the observed genetic effects along the chromosomes is in line with previous studies in wheat (Jordan et al. 2018) and barley (Casale et al. 2022) reporting that differences in the genome recombination rate among populations are explained by multiple loci with small effects.

A negative correlation was observed on a broad scale between recombination rate and the extent of methylation in the CpG and CHG sequence contexts (Figure 1B). This is in accordance with previous studies in other plant species showing that COs occurred in euchromatic regions while heterochromatic regions were depleted of COs and that hypomethylation at CpG sites increased the genome-wide CO rate (Melamed-Bessudo and Levy 2012; Colomé-Tatché et al. 2012; Salomé et al. 2012; Yelina et al. 2012; Mirouze et al. 2012; Wijnker et al. 2013; Rodgers-Melnick

et al. 2015; Marand et al. 2019; Apuli et al. 2020; Boideau et al. 2022; Fernandes et al. 2024). This association was confirmed at high-resolution by oberving the relationship between 10 kb coldspot windows and increased methylation in the CpG and CHG sequence contexts (Figure 3). Furthermore, compared with those in the nonhotspot windows, the methylation levels in the CpG and CHG sequence contexts in both the distal and pericentromeric regions of the barley chromosomes decreased in the hotspot windows. This result is in line with previous findings in maize showing a strong relationship between the occurrence of hotspots and decreased CpG and CHG methylation but no association with CHH methylation (Rodgers-Melnick et al. 2015). In this respect, it was suggested that increased recombination was associated with increased CHH methylation in regions with high CpG-related methylation levels but with decreased recombination where CpG methylation was low (Rodgers-Melnick et al. 2015).

The presence of SVs was shown to suppress COs in previous studies in *Arabidopsis* (Rowan et al. 2019; Fernandes et al. 2024) and other plant species (Rodgers-Melnick et al. 2015; Shen et al. 2019). In the present study, we were able to detect a decreased overlap and a longer distance between CO breakpoints and SVs compared to a random distribution, thus indicating the negative association of SVs with the occurrence of COs in barley (Tables S13 and S14). The type and size of the SVs were not found to be related to such effects, which is in line with previous findings in *Arabidopsis* (Rowan et al. 2019). Moreover, this is the first study showing the joint effect of methylation and the accumulation of SVs in determining genomic

regions deprived of recombination outside the pericentromeric region (Fernandes et al. 2024) and the variation in this effect within the genomic region (Figure 3 and Tables S11 and S12). In the distal telomeric region of the barley chromosomes, most 10 kb coldspot windows were found to be associated with increases in both the recombination level and the SV load. However, in the distal proximal region, increased methylation was found to be associated with most of the 10 kb coldspot windows, but an increased SV load was found to be associated with coldspot windows with no increased methylation. Interestingly, the effects of both methylation and SVs on the occurrence of coldspots were noticeable not only when comparing coldspot windows with other windows located far away in the same genomic region but also when comparing coldspots with their neighboring windows. This indicated a marked local effect of methylation and SVs on recombination suppression (Figure 4). In addition, the differences in both methylation level and SV load among barley populations were found to be responsible for the differences in the localization of coldspot windows among such populations. The parental inbred lines of the analyzed populations were found to differ in the methylation level in the genomic windows identified as coldspots in their offspring populations (Table 2), indicating that the inheritance of high methylation from only one parent was sufficient to prevent recombination in a particular region of the genome.

In a previous study on the same barley populations in which methylation was not separated from the genetic effects of genotypes, the effect of individual parental inbred lines was shown to be the major determinant of the recombination rate of

the respective biparental offspring populations (Casale et al. 2022). The increased methylation at genomic regions leading to coldspots might be an important part of the genetic effect of the parents on the recombination rate, which was negatively correlated with methylation in the present study.

In contrast to the above described association of methylation and SVs with the occurence of recombination coldspots, no significant differences between parental inbred lines were found in the methylation level of hotspot windows, indicating that parents must have equally low methylation at a genomic segment to allow a recombination hotspot. The employed window resolution of 10 kb, which is longer than the length reported earlier for recombination hotspots in other plant species (Choi and Henderson 2015), might be the reason for the lack of detection of SV effects on hotspots and the lack of differences in methylation between hotspots and their neighboring windows.

Our findings demonstrate that the recombination landscape in barley is highly predictable. Most of the recombination occurs in multiple short highly recombining sections in the distal regions of the chromosomes. These recombination hotspots are located in proximity to genes and where the levels of methylation and SV load are low enough to allow CO concretion. In this sense, such hotspots alternate with long regions deprived of recombination because of increased methylation or the accumulation of SVs preventing CO from occurring. Therefore, local differences in the recombination rate among barley populations can be explained to a considerable extent by differences in the methylation level and the accumulation of SVs at

multiple locations within the genome. Such differences are highly inheritable and can be determined by the effect of only one parent in a cross. However, our analyses suggest that in addition to these two genomic features, additional differences in the recombination machinery must exist, which forms the basis for what we designated genetic effect.

METHODS

Identification of the genomic features associated with recombination rate variation among barley populations

The recombination rates of 45 biparental barley populations, referred to as double round-robin populations, were obtained from Casale et al. (2022). These have been derived from genotyping the populations with the 50K Illumina Infinium iSelect SNP genotyping array (Bayer et al. 2017). The recombination rates were recalculated based on the Morex v3 reference genome sequence (Mascher et al. 2021) at 1 Mbp genomic windows. The pericentromeric region of each chromosome was defined as the continuous region surrounding the centromere for which the average recombination rate across the 45 DRR populations was 5-fold lower than the respective chromosome average across populations in 1 Mbp genomic windows. The regions of the chromosome that did not belong to the pericentromeric region were designated in the following as distal regions.

Whole-genome bisulfite DNA sequencing data for the 23 DRR parental inbred lines was obtained by extracting DNA from inbred lines from a mix of tissues, including the whole seedling plant, the leaf, and the apex, at stage 47 on the Zadoks scale. DNA library preparation was performed with NEBNext® Ultra™ II (New England Biolabs, Inc., USA), and bisulfite conversion was performed with the EZ DNA Methylation-Lightning Kit (Zymo Research, USA). The resulting 150 bp paired-end libraries were sequenced with Illumina HiSeq™ 2000 and NovaSeq™ (Il-

lumina, Inc., USA). The raw reads were trimmed with Trim Galore! (Krueger et al. 2023), mapped against the Morex v3 reference genome with Bismark (Krueger and Andrews 2011), and aligned with Bowtie 2 (Langmead and Salzberg 2012). For quality control, SNPs were called with Bis-SNP (Liu et al. 2012) and compared with single nucleotide variation (SNV) data generated by DNA sequencing of the respective inbred lines (Weisweiler et al. 2022). The level of methylation in cytosine positions present in the methylated sequence contexts CpG, CHG, and CHH, was calculated as the percentage of the methylated reads per position. For each DRR population, the methylation level at each sequence context in a given genomic window was calculated as the average among the respective parental inbred lines' methylation level values for the methylated cytosine positions in that window, weighted by the number of methylated cytosine positions corresponding to each parent. The average methylation level across the three sequence contexts in a given genomic 1 Mbp window was calculated as the average among the calculated methylation levels for such contexts in the window, weighted by the number of methylated cytosine positions corresponding to each of the contexts in the window. The difference in methylation level between the two parental inbred lines of a population at each 1 Mbp genomic window was calculated for the three methylated sequence contexts and their average.

The gene density in 1 Mbp windows was calculated as the physical fraction spanned by the coding sequence of genes in each window. The locations of genes and intergenic regions on the barley chromosomes were obtained from the Morex v3

reference sequence. The genetic divergence among parents of the DRR populations was calculated from single nucleotide variant (SNV) data derived from genome-wide sequencing (Weisweiler et al. 2022).

The SNVs were also used to calculate the general recombination effects (GRE) of the parental inbred lines as described by Casale et al. (2022). In the next step, the proportion of the sum of the GRE of both parents for a given population that was not explained by the average methylation in each 1 Mbp window was estimated using linear regression. This residual was assumed to represent the genetic effects on recombination in a given genomic window. The specific recombination effect (SRE) for a given parental combination was not taken into account to estimate genetic effects because it was previously described to cause only a minor effect on the recombination rate of a given biparental barley population (Casale et al. 2022).

The SVs such as inversions, insertions, deletions, duplications, and translocations, between the parental inbred lines of the DRR populations and the Morex reference genome were obtained from Weisweiler et al. (2022). The SVs were categorized by size (50—299 bp, 0.3—4.9 kb, 5—49 kb, 50—249 kb, 0.25—1 Mbp, and >1 Mbp), except for translocations whose length was not determined (Table S1). The physical length fraction spanned by SVs in every 1 Mbp genomic window across the genome was estimated for all SV categories and sizes. Furthermore, the total SV span fraction generated by the sum of all SV categories and sizes in each 1 Mbp window was calculated (hereafter referred to as SV load).

Pearson's correlation between all pairs of the abovementioned genomic features

was calculated by genomic window and population. To identify which of the features better explained the recombination rate variation among the 45 DRR populations at each 1 Mbp genomic window in the barley chromosomes, a stepwise regression approach was used. The procedure keeps for each 1 Mbp window the subset of genomic features that explain differences among the recombination rates of the DRR populations with the highest Akaike information criterion (AIC). The fraction of 1 Mbp windows of the barley genome in which a given genomic feature was retained in the model provides an estimation of the feature's importance across the entire genome. Moreover, the direction of the correlation between the analyzed features and the recombination rate provides a notion of the impact of the feature on either promoting or repressing recombination. The model included the total SV load, parental sequence divergence, genetic effects, average methylation level, and difference among parental inbred lines at the average methylation level. In addition, an extended model was constructed by breaking down the methylation-related variables into the respective methylated sequence contexts and the SV load into the different SV types.

Investigation of the genomic features associated with the recombination rate in barley at high-resolution

Plant material, genotyping, and data cleaning

From the 45 DRR population set, three populations (HvDRR13, HvDRR27, and HvDRR28) were selected for high-density genotyping using an mRNAseq approach

as described by Arlt et al. (2023). These populations are the product of a triangle cross among three parental inbred lines (HOR8160, SprattArcher, and Unumli-Arpa). The respective 64, 92, and 79 RILs were cultivated at the S7 generation in petri dishes in a randomized incomplete block design where the parental inbred lines were included as controls. The blocks were harvested 7 days after planting with less than 2 hours difference between the first and last sample. The whole seedling was utilized for mRNA extraction. For mRNA sequencing (RNA-Seq), the RIL-specific library was constructed using the VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina (Vazyme, China), and RIL-specific barcodes were used. The pooled libraries were sequenced on the DNBSEQ-G400 platform (MGI Tech Co., Ltd., China) by BGI Genomics (Beijing, China), generating 1.42 billion 150 bp paired-end reads. The reads were trimmed with Trimmomatic (Bolger et al. 2014) and aligned to the Morex v3 reference sequence using HISAT2 (Kim et al. 2019). Variant calling was performed using BCFtools (Li et al. 2009). The obtained variants were selected based on their intersection with the reported SNVs from the parental inbred lines (Weisweiler et al. 2022). Furthermore, a 12 K subset of the SNP markers reported previously (Casale et al. 2022) for the same parental inbred lines and RILs was added to the total set at genomic positions not present in the RNAseq dataset. On a population basis, SNPs carrying nonparental alleles were set to missing data, and SNPs with 100% missing data or monomorphic parental alleles were discarded. Missing data for genotypes at polymorphic positions were reconstructed using Beagle (Browning et al. 2018) with default parameters.

Detection of recombination events

A recombination event in a given RIL haplotype was called when a block of SNP alleles inherited from one parental inbred switched to a block of SNP alleles belonging to the other parent (i.e., parental allele phase change). The recombination breakpoints were determined as the midpoint of the region between both blocks (i.e., the CO interval). The blocks comprising fewer than three SNPs were considered false positive CO events and were discarded. Then, blocks shorter than 10 kb were considered GC events, while blocks longer than 10 kb were considered to be produced by CO (Yang et al. 2012; Gardiner et al. 2019). To enable comparisons with earlier studies (Yang et al. e.g. 2012; Gardiner et al. e.g. 2019), GC-related blocks were grouped into long (2–10 kb), short (20 bp–2 kb), and very short (2–19 bp) GC blocks, while the CO-related segments were grouped into short (10–500 kb), medium (500 kb–3 Mbp), and long (> 3 Mbp) CO blocks. The longest block length threshold that kept every major parental allele phase change (3 Mbp) was defined visually by graphical genotypes on a 500 kb scale from 0.5 to 20 Mbp (Figure S2). The CO block length categories were considered different CO layers, and further analyses were performed on a multilayer basis. Individuals with a CO count falling outside the 3-fold interquartile range of their respective population were assumed to be outliers and were discarded from further analyses.

The pericentromeric and distal regions of each chromosome were defined as explained above at 10 kb genomic windows for each of the three analyzed popula-

tions independently. The distal regions were defined as the chromosome segments between the pericentromeric region and the telomeres of the chromosomes.

In each 10 kb genomic window of the genome, the physical fraction spanned by the CO intervals determined in all RILs of a population was aggregated to calculate the accumulated CO probability per window on a population basis. The accumulated CO probabilities per window were normalized per chromosome and per population. The CO hotspot windows in a given population were defined as the windows with a CO probability $> 99\%$. The GC hotspot windows were determined in the same way as the CO hotspot windows. The windows located in the distal regions with a CO probability of zero were considered coldspot windows. The coldspot windows that were located beside other coldspot windows were combined into coldspot regions. To avoid calling coldspot windows near the pericentromeric region and telomeres, only windows located away from such regions were considered coldspot windows. This distance was granted by introducing an arbitrary margin with a length of 12.5% of the physical length of the respective distal region. To control for spurious associations generated by the variation of the features along the chromosomes, the windows from the distal regions were grouped into those close to the telomere (distal telomeric) and those close to the pericentromeric region (distal proximal).

Investigation of the association between genomic features and recombination rate

The fraction of each window spanned by SVs, the gene density, the methylation level at the sequence contexts CpG, CHG, and CHH, their average, and the parental difference for these contexts were calculated as described above for each 10 kb genomic window of the barley chromosomes in the three analyzed populations HvDRR13, HvDRR27, and HvDRR28. In addition, the 10 kb windows neighboring coldspot and hotspot regions (any genomic length spanned by contiguous coldspot or hotspot 10 kb windows) were grouped by their relative position in the range from -40 kb to +40 kb around the respective coldspot or hotspot in the mentioned chromosome regions. The statistical comparison among window groups of any kind for the mentioned features was performed with the Mann–Whitney U test with Bonferroni correction for multiple testing.

The observed overlaps of the coldspot and hotspot regions with genes and intergenic regions in the chromosomes of the three analyzed populations were statistically compared with random overlaps simulated with regioneR (Gel et al. 2016) by running $1,000$ permutations on a by-chromosome basis where the respective pericentromeric regions were masked.

The observed overlap between CO intervals and the genomic regions spanned by SVs was assessed as described above with regioneR for insertions/deletions, duplications, and inversions. In addition, the mean distance between the CO breakpoints

and their closest SV in the genome was compared with the equivalent of random simulated COs in 10 Mbp genomic windows, each with a probability of CO occurrence according to the recombination rate per chromosome in the DRR populations reported by Casale et al. (2022). The significant differences among the means of the observed and simulated CO-SV distances were evaluated with the Mann–Whitney U test.

DATA ACCESS

The methylation data of the 23 DRR population parental lines and the raw mRNA sequencing data of the RILs from populations HvDRR13, HvDRR27, and HvDRR28 (and their parental lines) are both available at NCBI, BioProject accessions PRJNA1100572 and 1088431, respectively. Employed scripts are available from the authors upon request.

COMPETING INTEREST STATEMENT

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. The authors give thanks to the IPK for providing the seeds of the diversity panel. We thank our former colleagues Andrea Lossow, Nicole Kliche-Kamphaus, Nele Kaul,

Isabelle Scheibert, Marianne Haperscheid, George Alskief, Florian Esser as well as our present colleagues Konstantin Shek and Stefanie Krey for their excellent technical assistance in creating and maintaining the DRR populations.

FUNDING

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111) and core funding of HHU and JKI.

AUTHORS' CONTRIBUTIONS

FC contributed to the design of the project, analysed the data and performed the analyses related to meiotic recombination, and performed all statistical analyses. CA created the libraries, sequenced, and processed the RNA-Seq data. MK analysed the bisulfite-sequenced DNA data. Both CA and MK contributes equally to this work as second authors. JL created the segregating populations. JE and TH created the bisulfite libraries for DNA sequencing. BS designed and coordinated the project. FC and BS wrote the manuscript. All authors read and approved the final manuscript.

Table 1: The observed overlaps among the recombination coldspot and hotspot 10 kb windows with genes and the intergenic regions in the distal region of the barley chromosomes, and their comparison with the overlaps generated under a random distribution of such regions in the analyzed double round-robin (DRR) populations HvDRR13, HvDRR27, and HvDRR28. The random distribution of the genomic regions was simulated with 1,000 permutations.

Recombination region	Genetic region	Population	Observed overlaps	Random overlaps
Coldspot windows	Genes	HvDRR13	23989	37029
		HvDRR27	19545	26369
		HvDRR28	24832	37091
	Intergenic regions	HvDRR13	77006	129894
		HvDRR27	73553	87290
		HvDRR28	78332	135060
Hotspot windows	Genes	HvDRR13	1057***	237
		HvDRR27	739***	146
		HvDRR28	1264***	282
	Intergenic regions	HvDRR13	823	840
		HvDRR27	684***	486
		HvDRR28	1057	1031

$^{***}P < 0.001$ for H_0: $\mu_{Observed} \leq \mu_{Random}$; Z-test.

Table 2: Comparison of the methylation level at sequence contexts CpG, CHG, and CHH in the windows identified as recombination hotspots and coldspots in the double round-robin (DRR) populations HvDRR13, HvDRR27, and HvDRR28 for the respective parental inbred lines. The coldspot and hotspots windows were calculated on the basis of the recombination rate of the DRR populations. For a given population and given methylated sequence context, the significant difference ($\alpha = 0.016$) in the Wilcoxon's rank sum test among genomic window groups are indicated with different letters.

	Coldspot windows				Hotspot windows			
Population	Offspring	HOR8160	Unumli-Arpa	SprattArcher	Offspring	HOR8160	Unumli-Arpa	SprattArcher
HvDRR13								
CpG	0.880 c	0.875 b	-	0.883 a	0.666 a	0.661 a	-	0.668 a
CHG	0.581 c	0.571 b	-	0.590 a	0.334 a	0.322 a	-	0.345 a
CHH	0.020 c	0.020 b	-	0.021 a	0.029 a	0.028 a	-	0.030 a
HvDRR27								
CpG	0.894 b	-	0.891 a	0.892 b	0.673 a	-	0.667 a	0.678 a
CHG	0.595 c	-	0.587 a	0.602 b	0.346 a	-	0.330 a	0.362 a
CHH	0.019 b	-	0.019 a	0.020 b	0.029 ab	-	0.027 a	0.031 b
HvDRR28								
CpG	0.880 c	0.875 b	0.881 a	-	0.645 a	0.643 a	0.644 a	-
CHG	0.573 c	0.570 b	0.574 a	-	0.319 a	0.319 a	0.315 a	-
CHH	0.019 a	0.019 b	0.019 a	-	0.028 a	0.028 a	0.027 a	-

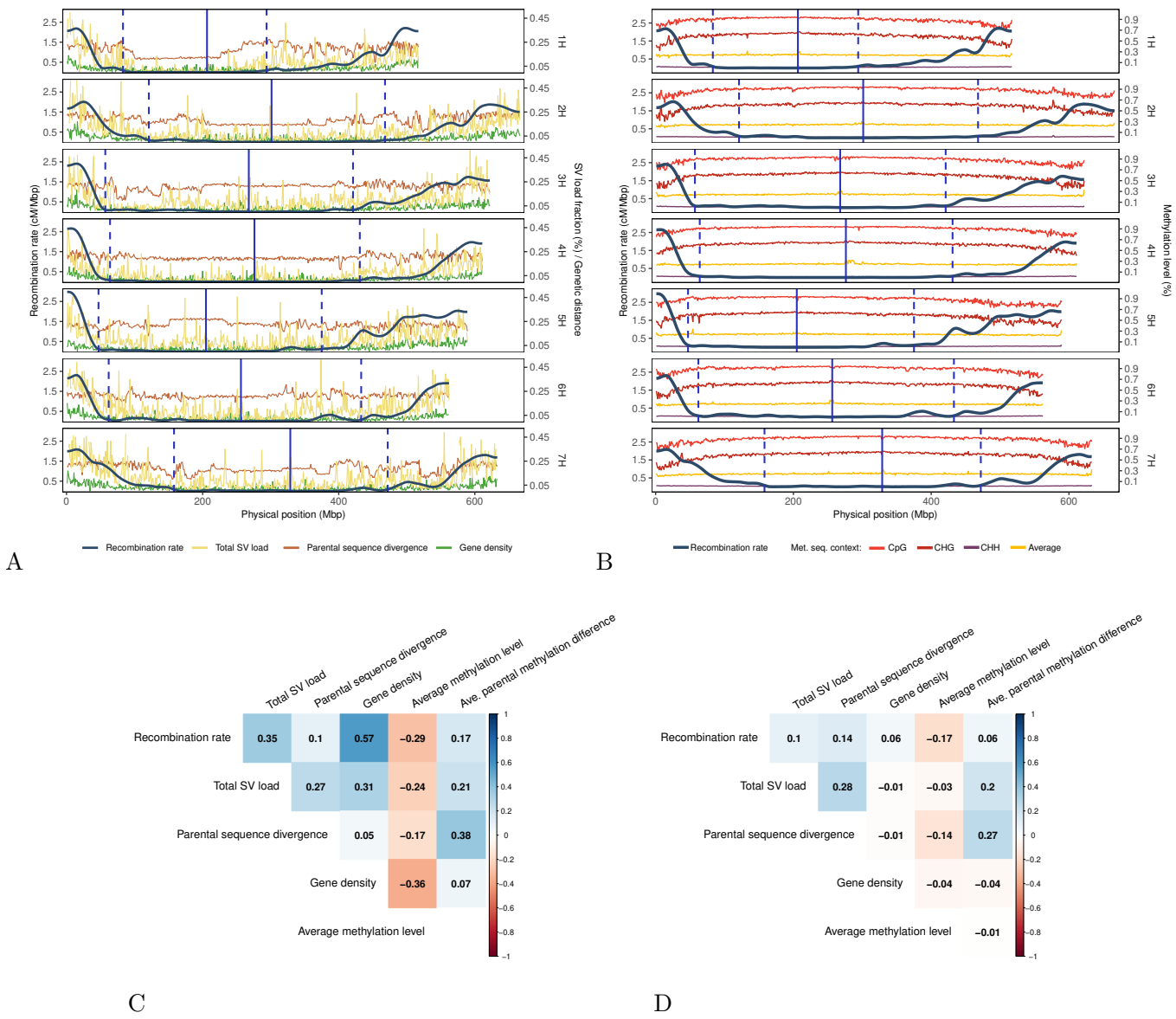
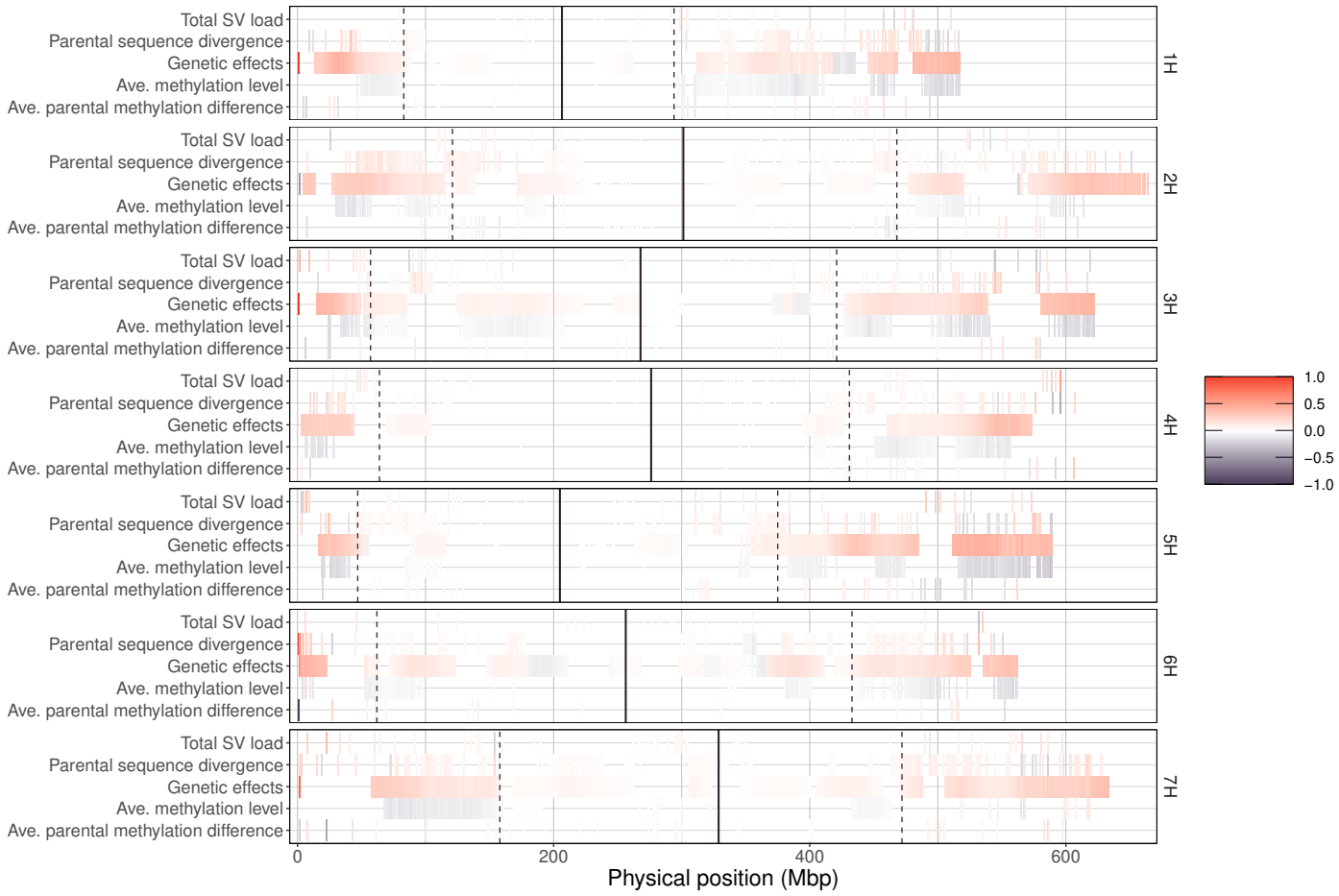


Fig. 1: (A) Distribution of recombination rate, total structural variants' load fraction, gene density, and parental sequence divergence between the respective parental inbred lines on average across the 45 double round-robin (DRR) populations in 1 Mbp windows across the seven barley chromosomes. The total structural variants (SV) load fraction represents the portion spanned of a given 1 Mbp genomic window by the sum of insertions, deletions, invertions, and duplications. (B) Distribution of recombination rate and methylation level in 1 Mbp windows for the sequence contexts CpG, CHG, CHH, and their mean, on average across the seven barley chromosomes and 45 DRR populations. The methylation level values for a given methylation sequence context represent the percentage of methylated reads of such context present in a 1 Mbp window averaged across the 45 DRR populations. The vertical blue solid line indicates the position of the centromere in the Morex v3 reference genome and the vertical blue dashed lines indicate the pericentromeric region calculated across the 45 DRR populations. (C-D) Correlation matrix between recombination rate, total SV load fraction, parental sequence divergence, gene density, average methylation level, and parental difference in methylation level, across 1 Mbp windows in the distal (C) and the pericentromeric (D) regions of the barley chromosomes for the average across the 45 DRR populations. Pearson's correlation coefficients are indicated with a color gradient from -1 (red) to 1 (blue).



Genomic feature	Chromosome region			
	Distal		Pericentromeric	
	+	-	+	-
Total SV load	0.05	0.02	0.04	0.01
Parental sequence divergence	0.24	0.04	0.19	0.01
Genetic effects	0.76	0.01	0.40	0.06
Average methylation level	0.00	0.43	0.01	0.12
Ave. parental methylation difference	0.04	0.03	0.02	0.02

Fig. 2: The genomic features explaining the variation in the recombination rate among the 45 double round-robin (DRR) populations in 1 Mbp windows across the seven barley chromosomes. The genomic features were selected for each window by a stepwise regression procedure. The standarized regression coefficients are indicated by a color gradient from -1 (purple) to 1 (red). The studied genomic features included sequence divergence among parental inbred lines, genetic effects, total structural variants (SV) load, average methylation level across the sequence contexts CpG, CHG, and CHH, and the difference in the methylation level among the parental inbred lines for the average of such sequence contexts. The vertical solid line indicates the position of the centromere in the reference genome, and the vertical dashed lines indicate the pericentromeric region calculated across the 45 DRR populations. The analysis was performed separately for the distal and pericentromeric regions of the barley chromosomes, respectively. The fraction of 1 Mbp windows of the barley genome in which the genomic features were found to be significantly associated with the recombination rate variation across the 45 DRR populations are displayed in the table. The fractions of 1 Mbp windows positively and negatively associated with the recombination rate were indicated with the "+" or the "−" signs, respectvelv.

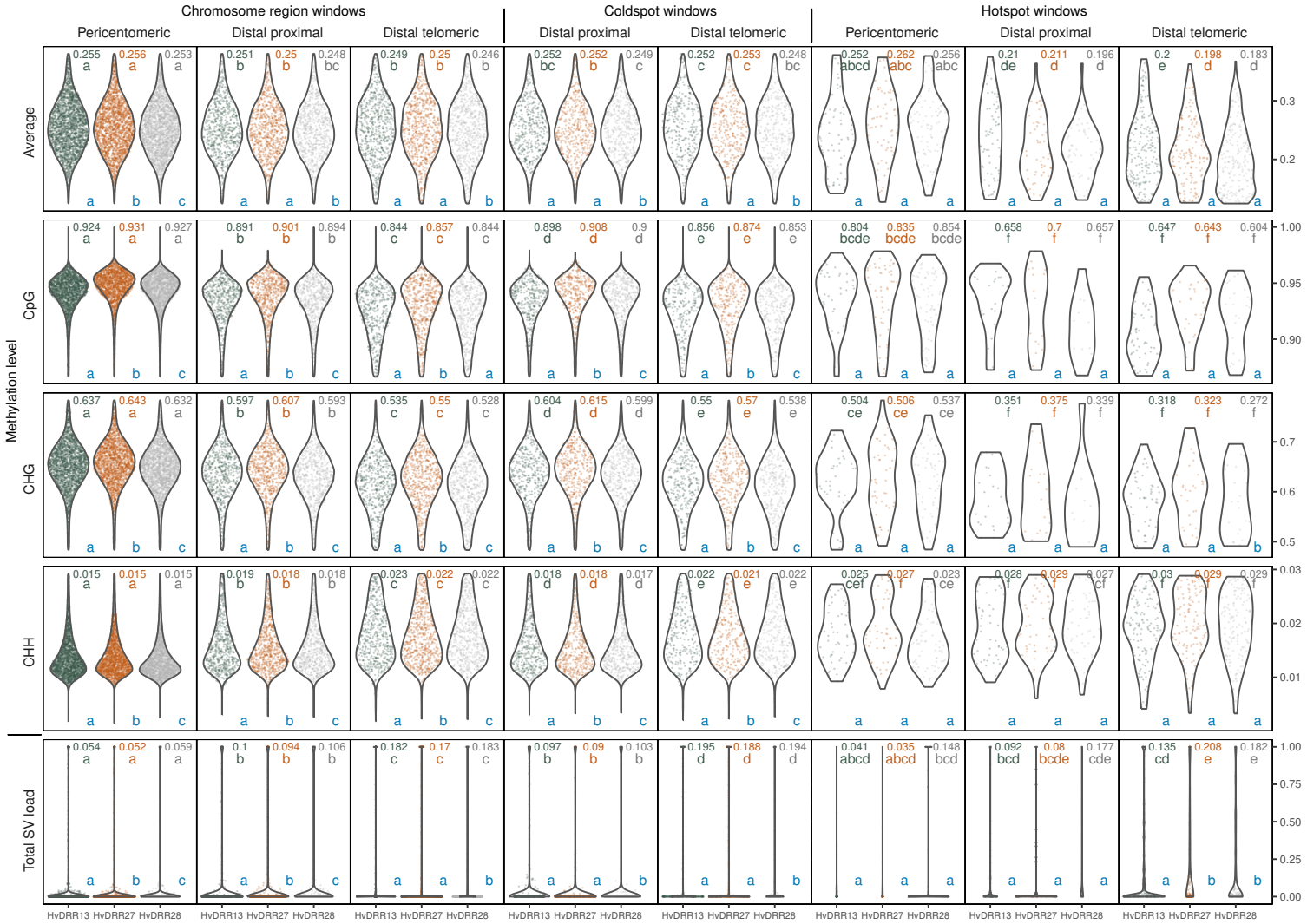


Fig. 3: Distribution of the methylation level for the methylated sequence contexts CpG, CHG, CHH, and their average, and the total structural variants (SV) load fraction in 10 kb genomic windows grouped by their location in different chromosomal regions -pericentromeric, distal proximal, and distal telomeric-, and recombination rate -total, coldspots, and hotspots- of the barley chromosomes in the analyzed double round-robin (DRR) populations HvDRR13 (green), HvDRR27 (orange), and HvDRR28 (grey). For each DRR population, the methylation level at each sequence context in a given genomic window was calculated as the average among the respective parental inbred lines' methylation level values for the methylated cytosine positions in that window, weighted by the number of methylated cytosine positions corresponding to each parent. The displayed dots for chromosome regions and coldspots are a random subset of 1% of the total windows in each window group. For each genomic feature, the distribution's mean of each window group from a given population is indicated with the related population color at the top-right of the respective plot. Significant differences ($\alpha = 0.001$) in the Wilcoxon's rank sum test among such means across window groups is indicated with different letters below the respective mean and sharing the population color. For a given genomic feature in each window grouping category, significant differences ($\alpha = 0.008$) in the Wilcoxon's rank sum test among populations are indicated with different blue letters at the bottom of each panel.

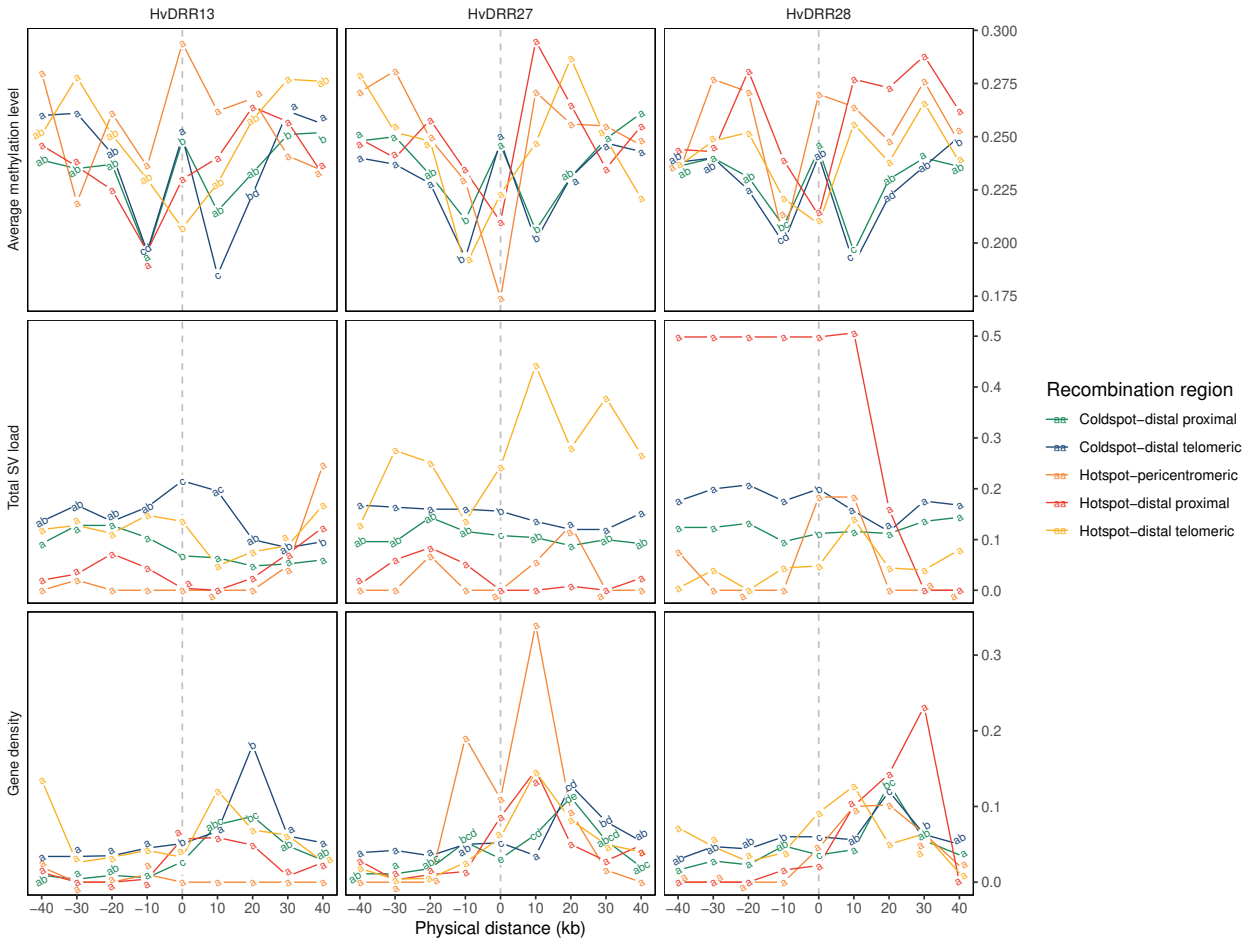


Fig. 4: The mean values of the average methylation level among the methylated sequence contexts CpG, CHG, and CHH, total structural variants (SV) load, and gene density in the genomic windows grouped by the physical positions in the range from -40 kb to +40 kb around the coldspot and hotspot genomic regions in the pericentromeric, distal proximal, and distal telomeric region of the chromosomes in the analyzed double round-robin (DRR) populations HvDRR13, HvDRR27, and HvDRR28. The vertical dashed lines indicate the relative location of either the coldspot or the hotspot windows, respectively. The significant difference ($\alpha = 0.001$) in the Wilcoxon's rank sum test among the window groups corresponding to the different 10 kb physical position neighboring the respective hotspot or coldspot region in a particular chromosome region is indicated with different letters.

REFERENCES

Apuli RP, Bernhardsson C, Schiffthaler B, Robinson KM, Jansson S, Street NR, and Ingvarsson PK. 2020. Inferring the genomic landscape of recombination rate variation in european aspen (Populus tremula). *G3: Genes, Genomes, Genetics* **10**: 299–309.

Arbeithuber B, Betancourt AJ, Ebner T, and Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences* **112**: 2109–2114.

Arlt C, Wachtmeister T, Köhrer K, and Stich B. 2023. Affordable, accurate and unbiased RNA sequencing by manual library miniaturization: A case study in barley. *Plant Biotechnology Journal* **21**: 2241–2253.

Barton NH and Charlesworth B. 1998. Why sex and recombination? *Science* **281**: 1986–1990.

Baudat F and Massy BD. 2007. Regulating double-stranded dna break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome research* **15**: 565–577.

Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W, et al.. 2013. Intraspecific variation of recombination rate in maize. *Genome Biology* **14**.

Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, et al.. 2017. Development and evaluation of a barley 50k iselect SNP array. *Frontiers in Plant Science* **8**: 1792.

Blackwell AR, Dluzewska J, Szymanska-Lejman M, Desjardins S, Tock AJ, Kbiri N, Lambing C, Lawrence EJ, Bieluszewski T, Rowan B, et al.. 2020. MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. *The EMBO Journal* **39**: 1–22.

Blary A and Jenczewski E. 2019. Manipulation of crossover frequency and distribution for plant breeding. *Theoretical and Applied Genetics* **132**: 575–592.

Boideau F, Richard G, Coriton O, Huteau V, Belser C, Deniot G, Eber F, Falentin C, de Carvalho JF, Gilet M, et al.. 2022. Epigenomic and structural events preclude recombination in Brassica napus. *New Phytologist* **234**: 545–559.

Bolger AM, Lohse M, and Usadel B. 2014. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bouchet S, Olatoye MO, Marla SR, Perumal R, Tesso T, Yu J, Tuinstra M, and Morris GP. 2017. Increased Power To Dissect Adaptive Traits in Global Sorghum

Diversity Using a Nested Association Mapping Population. *Genetics* **206**: 573–585.

Browning BL, Zhou Y, and Browning SR. 2018. A one-penny imputed genome from next-generation reference panels. *American Journal of Human Genetics* **103**: 338–348.

Burt A. 2000. Perspective: Sex, recombination, and the efficacy of selection - was weismann right? *Evolution* **54**: 337–351.

Casale F, Van Inghelandt D, Weisweiler M, Li J, and Stich B. 2022. Genomic prediction of the recombination rate variation in barley – A route to highly recombinogenic genotypes. *Plant Biotechnology Journal* **20**: 676–690.

Choi K and Henderson IR. 2015. Meiotic recombination hotspots - a comparative view. *Plant Journal* **83**: 52–61.

Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FCH, et al.. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics* **45**: 1327–1338.

Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouz B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, et al.. 2012. Features of the arabidopsis recombination landscape resulting from the combined loss of sequence variation and dna methylation. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 16240–16245.

Darrier B, Rimbert H, Balfourier F, Pingault L, Josselin AA, Servin B, Navarro J, Choulet F, Paux E, and Sourdille P. 2017. High-resolution mapping of crossover events in the hexaploid wheat genome suggests a universal recombination mechanism. *Genetics* **206**: 1373–1388.

Dreissig S, Mascher M, Heckmann S, and Purugganan M. 2019. Variation in recombination rate is shaped by domestication and environmental conditions in barley. *Molecular Biology and Evolution* **36**: 2029–2039.

Dreissig S, Maurer A, Sharma R, Milne L, Flavell AJ, Schmutzer T, and Pillen K. 2020. Natural variation in meiotic recombination rate shapes introgression patterns in intraspecific hybrids between wild and domesticated barley. *New Phytologist* **228**: 1852–1863.

Esch E, Szymaniak JM, Yates H, Pawlowski WP, and Buckler ES. 2007. Using crossover breakpoints in recombinant inbred lines to identify quantitative trait loci controlling the global recombination frequency. *Genetics* **177**: 1851–1858.

Fernandes JB, Naish M, Lian Q, Burns R, Tock AJ, Rabanal FA, Wlodzimierz P, Habring A, Nicholas RE, Weigel D, et al.. 2024. Structural variation and dna methylation shape the centromere-proximal meiotic crossover landscape in arabidopsis. *Genome Biology* **25**: 30.

Gardiner LJ, Wingen LU, Bailey P, Joynson R, Brabbs T, Wright J, Higgins JD, Hall N, Griffiths S, Clavijo BJ, et al.. 2019. Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency. *Genome Biology* **20**: 69.

Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, and Malinverni R. 2016. RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289–291.

Gion JM, Hudson CJ, Lesur I, Vaillancourt RE, Potts BM, and Freeman JS. 2016. Genome-wide variation in recombination rate in Eucalyptus. *BMC Genomics* **17**: 1–12.

Giraut L, Falque M, Drouaud J, Pereira L, Martin OC, and Mézard C. 2011. Genome-wide crossover distribution in arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics* **7**: e1002354.

Gutierrez-Gonzalez JJ, Mascher M, Poland J, and Muehlbauer GJ. 2019. Dense genotyping-by-sequencing linkage maps of two Synthetic W7984×Opata reference populations provide insights into wheat structural diversity. *Scientific Reports* **9**: 1–15.

Hall JC. 1972. Chromosome segregation influenced by two alleles of the meiotic mutant c(3)G in Drosophila melanogaster. *Genetics* **71**: 367–400.

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, et al.. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**.

Henderson IR. 2012. Control of meiotic recombination frequency in plant genomes. *Current Opinion in Plant Biology* **15**: 556–561.

Higgins JD, Perry RM, Barakate A, Ramsay L, Waugh R, Halpin C, Armstrong SJ, and Franklin FCH. 2012. Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in barley. *Plant Cell* **24**: 4096–4109.

Hsu YM, Falque M, and Martin OC. 2022. Quantitative modelling of fine-scale variations in the Arabidopsis thaliana crossover landscape. *Quantitative Plant Biology* **3**: e3.

Jordan KW, Wang S, He F, Chao S, Lun Y, Paux E, Sourdille P, Sherman J, Akhunova A, Blake NK, et al.. 2018. The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *Plant Journal* **95**: 1039–1054.

Kim D, Paggi JM, Park C, Bennett C, and Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**: 907–915.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, and Nordborg M. 2007. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics* **39**: 1151–1155.

Krueger F and Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**: 1571–1572.

Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, Hulselmans G, and sclamons. 2023. Trimgalore v0.6.10.

Langmead B and Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357–359.

Lawrence EJ, Griffin CH, and Henderson IR. 2017. Modification of meiotic recombination by natural variation in plants. *Journal of Experimental Botany* **68**: 5471–5483.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li X, Li L, and Yan J. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature Communications* **6**.

Liu Y, Siegmund K, Laird P, and Berman B. 2012. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. *Genome biology* **13**: R61.

Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T, and Ma H. 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing landsberg erecta and all four products of a single meiosis. *Genome Research* **22**: 508–518.

Mancera E, Bourgon R, Brozzi A, Huber W, and Steinmetz M. 2009. High-resolution mapping of meiotic crossovers and noncrossovers in yeast. *Nature* **454**: 479–485.

Marand AP, Zhao H, Zhang W, Zeng Z, Fang C, and Jianga J. 2019. Historical meiotic crossover hotspots fueled patterns of evolutionary divergence in rice. *Plant Cell* **31**: 645–662.

Martini E, Diaz RL, Hunter N, and Keeney S. 2006. Crossover homeostasis in yeast meiosis. *Cell* **126**: 285–295.

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, Gundlach H, Boston LB, Tulpová Z, et al.. 2021. Long-read sequence assembly: A technical evaluation in barley. *Plant Cell* **33**: 1888–1906.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, et al.. 2009. Genetic properties of the maize nested association mapping population. *Science* **325**: 737–740.

Melamed-Bessudo C and Levy AA. 2012. Deficiency in dna methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **109**.

Mercier R, Mézard C, Jenczewski E, Macaisne N, and Grelon M. 2015. The molecular biology of meiosis in plants. *Annual Review of Plant Biology* **66**: 297–327.

Mirouze M, Lieberman-Lazarovich M, Aversano R, Bucher E, Nicolet J, Reinders J, and Paszkowski J. 2012. Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 5880–5885.

Muller HJ. 1932. Some genetic aspects of sex. *The American Naturalist* **66**: 118–138.

Mézard C. 2006. Meiotic recombination hotspots in plants. *Biochemical Society Transactions* **34**: 531–534.

Nachman MW. 2002. Variation in recombination rate across the genome: Evidence and implications. *Current Opinion in Genetics and Development* **12**: 657–663.

Paape T, Zhou P, Branca A, Briskine R, Young N, and Tiffin P. 2012. Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. *Genome Biology and Evolution* **4**: 726–737.

Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, et al.. 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144**: 719–731.

Peck JR. 1994. A ruby in the rubbish: Beneficial mutations, deleterious mutations and the evolution of sex. *Genetics* **137**: 597–606.

Qi J, Chen Y, Copenhaver GP, and Ma H. 2014. Detection of genomic variations and dna polymorphisms and impact on analysis of meiotic recombination and

genetic mapping. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 10007–10012.

Ritz KR, Noor MA, and Singh ND. 2017. Variation in recombination rate: Adaptive or not? *Trends in Genetics* **33**: 364–374.

Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li C, Li Y, and Buckler ES. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 3823–3828.

Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, and Weigel D. 2019. An ultra high-density Arabidopsis thaliana crossover. *Genetics* **213**: 771–787.

Saintenac C, Faure S, Remay A, Choulet F, Ravel C, Paux E, Balfourier F, Feuillet C, and Sourdille P. 2011. Variation in crossover rates across a 3-mb contig of bread wheat (triticum aestivum) reveals the presence of a meiotic recombination hotspot. *Chromosoma* **120**: 185–198.

Salomé PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, and Weigel D. 2012. The recombination landscape in Arabidopsis thaliana f2 populations. *Heredity* **108**: 447–455.

Serra H, Choi K, Zhao X, Blackwell AR, Kim J, and Henderson IR. 2018. Interhomolog polymorphism shapes meiotic crossover within the Arabidopsis RAC1 and RPP13 disease resistance genes. *PLoS Genetics* **14**.

Shen C, Wang N, Huang C, Wang M, Zhang X, and Lin Z. 2019. Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. *Plant Journal* **99**: 494–505.

Si W, Yuan Y, Huang J, Zhang X, Zhang Y, Zhang Y, Tian D, Wang C, Yang Y, and Yang S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytologist* **206**: 1491–1502.

Silva-Junior OB and Grattapaglia D. 2015. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. *New Phytologist* **208**: 830–845.

Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, Michelmore RW, Huettel B, and Schneeberger K. 2019. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nature Communications* **10**: 1–9.

Sun Y, Ambrose JH, Haughey BS, Webster TD, Pierrie SN, Muñoz DF, Wellman EC, Cherian S, Lewis SM, Berchowitz LE, et al.. 2012. Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in Arabidopsis thaliana. *PLoS Genetics* **8**: e1002968.

Szostak JW, Orr-Weaver TL, Rothstein RJ, and Stahl FW. 1983. The double-strand-break repair model for recombination. *Cell* **33**: 25–35.

Wang J, Street NR, Scofield DG, and Ingvarsson PK. 2016. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related populus species. *Genetics* **202**: 1185–1200.

Weisweiler M, Arlt C, Wu PY, Inghelandt DV, Hartwig T, and Stich B. 2022. Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. *Theoretical and Applied Genetics* **135**: 3511–3529.

Wijnker E, James GV, Ding J, Becker F, Klasen JR, Rawat V, Rowan BA, de Jong DF, de Snoo CB, Zapata L, et al.. 2013. The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *eLife* **2**: e01426.

Yang S, Yuan Y, Wang L, Li J, Wang W, Liu H, Chen JQ, Hurst LD, and Tian D. 2012. Great majority of recombination events in Arabidopsis are gene conversion events. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 20992–20997.

Yelina NE, Choi K, Chelysheva L, Macaulay M, de Snoo B, Wijnker E, Miller N, Drouaud J, Grelon M, Copenhaver GP, et al.. 2012. Epigenetic remodeling of meiotic crossover frequency in arabidopsis thaliana DNA methyltransferase mutants. *PLoS Genetics* **8**: e1002844.

7 List of publications

1. Weisweiler, M., **Arlt, C.***, Wu, P.-Y.*, van Inghelandt, D., Hartwig, T., and Stich, B. (2022). Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. Theoretical and Applied Genetics, 135:3511–3529.

2. **Arlt, C.**, Wachtmeister, T., Köhrer, K., Stich, B. (2023). Affordable, accurate and unbiased RNA sequencing by manual library miniaturization: A case study in barley. Plant Biotechnology Journal, 21(11).

3. **Arlt, C.**, van Inghelandt, D., Li, J., Stich, B. (2025). Assessment of genomic prediction capabilities of transcriptome data in a barley multi-parent RIL population. Theoretical and Applied Genetics, Accepted.

4. Casale, F., **Arlt, C.**, Kühl, M., Li, J., Engelhorn, J., Hartwig T., Stich, B. (2025). The role of methylation and structural variants in shaping the recombination landscape of barley. Submitted for publication.

5. Rouina, H., Singh, D., **Arlt, C.**, Malekian, B., Schreiber, L., Stich, B., Marshall-Colon, A., (2025). Regulatory Analysis of Root Architectural and Anatomical Adaptation to Nitrate and Ammonium in Brachypodium distachyon. Submitted for publication.

*Contributed equally

8 Acknowledgments

I am very grateful to my academic supervisor: Prof. Dr. Benjamin Stich for his advice, suggestions, and support during this thesis work.

Many thanks to Prof. Dr. Benjamin Stich, Prof. Dr. Karl Köhrer, Dr. Thorsten Wachtmeister, Dr. Delphine van Inghelandt, Dr. Jinquan Li for being my co-authors of my first author's publications.

Special thanks to Stephanie Krey for her continued support in the laboratory and the hard work that made this research possible.

Many thanks to Prof. Dr. Karl Köhrer, Dr. Thorsten Wachtmeister and the Staff of the Genomics & Transcriptomics Laboratory (GTL) for the support.

Thanks to Prof. Dr. Benjamin Stich, Annett Sitte, Ines Sigge, Dr. Maria Schmidt, Dr. Bernd Hackauf, Dr. Suresh Bontha, Dr. Marius Weisweiler, Dr. Federico Casale, Dr. Po-Ya Wu, Dr. Asis Shrestha, Dr. Delphine van Inghelandt, Nadia Baig, Marius Kühl, Dr. Michael Schneider, Francesco Cosenza, Ricardo Guerreiro, Yanrong Gao, Kefas Luka Baiyi, Alessio Maggiorelli, Anjali Walpola Mudalige Dona, Twinkal Lapasiya, Aashu, Kathrin Thelen, Amelie Kok, Konstantin Shek, and all unmentioned members and former members of the Institute for Quantitative Genetics and Genomics of Plants and the JKI at Groß Lüsewitz for creating a great and pleasant work atmosphere.

Special thanks to my family for always encouraging, supporting, and believing in me.

-

Besonderen Dank an meine Familie für die stetige Unterstützung, Ermutigung und dafür, dass sie immer an mich geglaubt haben.