Heinrich Heine Universität Düsseldorf

Next generation sequencing approaches to facilitate the breeding of barley and potato

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Marius Weisweiler

aus Eschweiler

Düsseldorf, März 2022

aus dem Institut für Quantitative Genetik
und Genomik der Pflanzen
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

Berichterstatter:
1. Prof. Dr. B. Stich
2. Prof. Dr. G. Klau

Tag der mündlichen Prüfung: 28.09.2022

Eidesstattliche Versicherung und Selbstständigkeitserklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität" erstellt worden ist.
Außerdem versichere ich, dass ich diese Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren kein gescheitertes Promotionsverfahren vorausgegangen ist.

Düsseldorf, den 30.03.2022

_____

Marius Weisweiler

Table of contents

---

[1] Freire*, R., M. Weisweiler*, R. Guerreiro*, N. Baig, B. Hüttel, E. Obeng-Hinneh, J. Renner, St. Hartje, K. Muders, B. Truberg, A. Rosen, V. Prigge, J. Bruckmüller, J. Lübeck, B. Stich. 2021. G3 Genes|Genomes|Genetics 11:jkab330

[2] Weisweiler*, M., A. de Montaigu*, D. Ries, M. Pfeifer, B. Stich. 2019. BMC Genomics 20: 787

[3] Weisweiler, M., B. Stich. 2022. In preparation

[4] Weisweiler, M., C. Arlt*, P.-Y. Wu*, D. Van Inghelandt, T. Hartwig, B. Stich. 2022. PLOS Genetics In review

*Contributed equally

1. Summary

Improvements in breeding were responsible for crop productivity increase over the last century. Nevertheless, it is nowadays still crucial to further increase the crop productivity to meet the demands of the growing world population. This likewise requires proceeding improvements in breeding, e.g. a better understanding of the genetic background information to accelerate breeding cycles as well as to develop improved crop varieties. Therefore, a further characterization of genomic variation including the examination of new genetic markers can lead to an improvement in breeding programs.

The detection of various sequence variants used as genetic markers in complex crop genomes became considerably simpler due to the development of next generation sequencing (NGS). However, for the detection of sequence variants based on NGS, high quality reference sequences need to be available. I therefore used a combination of NGS technologies to create a high quality reference sequence for a potato clone derived from an elite variety and observed a high divergence between the new reference sequence and that of other potato varieties. Additionally, I evaluated the usefulness of this new reference sequence for various genomic approaches and illustrated the high potential to use it for breeding applications.

A genomic layer which was suggested to contribute to phenotypic variation of agronomic importance in barley was presence/absence variation (PAV). Thus, a multi-tissue mRNA sequencing approach was used to examine the genomic and expression PAV, as well as the transcriptomic variation in 23 spring barley inbreds which were selected from a panel of a world-wide collection and which are the parents of a new resource for joint linkage and association mapping, the double round robin population. Examining expression PAV data, I observed that more than 50% of all genes were not expressed in the 23 barley inbreds. Furthermore, an approach was developed to detect genomic PAV based on the identification of expression PAV.

Because of these promising results, the interest arose to systematically explore structural variants (SV) in the barley and potato genomes. For both crops, only single SV have been identified and associated with qualitative phenotypic traits. I therefore evaluated different SV detection algorithms using computer simulations on short-read and linked-read sequencing considering various technical and genomic features as the SV type, SV length, as well as the sequencing

coverage to find the best combination of algorithms and sequencing approaches to detect SV in the barley and potato genomes. Based on the simulation scenarios, the best combination of SV callers reached a sensitivity $> 70\%$ and precision $> 95\%$ for all combinations of SV types and SV lengths in the barley genome. Even higher sensitivity and precision values were observed using linked-read sequencing in the potato genome. Additionally, the simulation scenarios in the potato genome illustrated the respective strengths of linked- and short-read sequencing signals for SV detection, which are the determination of the SV length of large insertions due to the usage of long-range molecule signals and the detection of short SV by considering short-read signals.

The optimal combination of SV callers were then used to study the occurrence and distribution of SV in the barley genome. The SV detected by a DNA sequencing approach were associated with genome-wide gene and gene-specific gene expression. Further, SV, expression PAV, and gene expression data detected in barley showed additional genetic information compared to single nucleotide variants (SNV) and were therefore used to predict different phenotypic traits and showed an increased prediction ability compared to the classical approach using a SNV array. These findings suggest the usefulness of exploiting SV information when fine mapping and cloning the causal gene underlying quantitative traits as well as the high potential of using SV, expression PAV, and gene expression data for the prediction of phenotypes in diverse germplasm sets.

2. Zusammenfassung

Fortschritte in der Züchtung waren für eine Produktionssteigerung von Nutzpflanzen während des letzten Jahrhunderts verantwortlich. Dennoch ist es auch in der heutigen Zeit notwendig eine weitere Produktionssteigerung von Nutzpflanzen zu erzielen, um die Bedürfnisse einer wachsenden Weltbevölkerung zu decken. Ein Teil dieser Produktionssteigerung muss auch heutzutage durch Züchtung erzielt werden, beispielsweise durch ein besseres Verständnis der genetischen Information von Nutzpflanzen, wodurch Züchtungszyklen beschleunigt werden können und verbesserte Nutzpflanzensorten gezüchtet werden können. Deshalb kann eine fortschreitende Charakterisierung der genomischen Variation einschließlich der Analyse von neuen genetischen Markern zu einem Fortschritt in Züchtungsprogrammen führen.

Mit der Erfindung der Hochdurchsatzsequenzierung wurde die Identifikation von verschiedenen Sequenzvarianten, welche als genetische Marker in Nutzpflanzen benutzt werden, wesentlich erleichtert. Allerdings werden für die Identifizierung von Sequenzvarianten basierend auf der Hochdurchsatzsequenzierung hoch qualitative Referenzsequenzen benötigt. Deshalb nutzte ich eine Kombination aus Hochdurchsatzsequenzierungstechnologien, um eine hoch qualitative Referenzsequenz des Kartoffelgenoms einer Züchtungselitesorte zu erstellen und beobachtete eine hohe Diversität zwischen der neuen Referenz und denen der bereits vorhandenen Kartoffelsorten. Die Nutzung dieser neuen Referenzsequenz wurde für verschiedene genomische Methoden evaluiert, wodurch sich ein hohes Potential zur Nutzung in züchtungs-relevanten Anwendungen zeigte.

Eine genomische Eigenschaft, die zu phänotypischer Variation von agronomischer Bedeutung in Gerste beitragen soll, sind Anwesenheit/Abwesenheit Variationen (PAV). Daher wurde mRNA-Sequenzierung verschiedener Gewebe verwendet, um genomische und transkriptomische PAV, sowie trankriptomische Variation in 23 Sommergersten-Inzuchtlinien, die aus einer weltweiten Sammlung stammen und die die Eltern einer neuen Ressource für Kopplungs- und Assoziationskartierung sind, zu untersuchen. Durch die Analyse der transkriptomischen PAV zeigte sich, dass mehr als 50% aller Gene nicht in allen 23 Gersten-Inzuchtlinien expremiert waren. Des Weiteren wurde eine Methode entwickelt, wie trankriptomische PAV basierend auf der Identifizierung von genomischen PAV ermittelt werden können.

Aufgrund dieser vielversprechenden Ergebnisse entstand ein Interesse, strukturelle Varianten (SV) im Gersten- und Kartoffelgenom systematisch zu untersuchen. In Gerste und Kartoffel wurden bisher nur einzelne SV mit qualitativen phänotypischen Merkmalen assoziiert. Deshalb evaluierte ich verschiedene Algorithmen, die SV identifizieren, basierend auf Computer-Simulationen und zwei verschiedenen Ansätzen der Hochdurchsatzsequenzierung, unter der Berücksichtigung von verschiedenen technischen und genomischen Charakteristika wie den SV-Typ, die SV-Länge und die Sequenziertiefe, um die beste Kombination aus Algorithmen und Sequenziertechniken zu finden, um SV im Gersten- und Kartoffelgenom zu bestimmen. Anhand der Simulationen wurde gezeigt, dass die beste Kombination aus Algorithmen eine Sensitivität $> 70\%$ und eine Präzision $> 95\%$ für alle Kombinationen von SV-Typen und SV-Längen im Gerstengenom aufwies. Im Kartoffelgenom wurden noch höhere Sensitivitäts- und Präzisionswerte beobachtet. Des Weiteren zeigten die Simulationen im Kartoffelgenom die jeweiligen Stärken zweier Hochdurchsatzsequenzierungsansätze zur Bestimmung von SV, welche die Bestimmung der SV-Länge langer Insertionen durch *linked-read* Sequenzierung und die Identifizierung kurzer SV durch *short-read* Signale sind.

Die optimale Kombination aus Algorithmen wurde dann verwendet, um das Auftauchen und die Verteilung von SV im Gerstengenom zu untersuchen. DNA-Sequenzierung wurde genutzt, um SV zu identifizieren, welche mit genom-weiter und gen-spezifischer Genexpression assoziiert wurden. Des Weiteren zeigte sich, dass transkriptomische PAV, SV und Genexpressionsdaten, die in den Gersten-Inzuchtlinien bestimmt wurden, zusätzliche genetische Informationen verglichen zu Einzelnukleotid-Varianten (SNV) beinhalten. Deshalb wurden diese für die Vorhersage von verschiedenen phänotypischen Merkmalen verwendet, bei der eine höhere Vorhersage-Genauigkeit verglichen zu der klassischen Nutzung eines SNV *arrays* aufgewiesen wurde. Diese Ergebnisse zeigen sowohl den Nutzen, SV Informationen zu berücksichtigen, wenn Genloci kodierend für quantitative Merkmale identifiziert wurden und das zugrundeliegende Gen analysiert wird, als auch das große Potential, SV, transkriptomische PAV und Genexpressionsdaten für die Vorhersage der Phänotypen in Datensätzen mit hoher genetischer Vielfalt zu nutzen.

3. Introduction

To meet the demands of the growing world population, one of the most important issues of modern agriculture is to increase the productivity of crops (Frona et al., 2019). The limitation of cultivated land and water (Beddington et al., 2012) makes it crucial to increase the yield of crops. Additionally, due to the climate change and corresponding increased temperatures, crops in breeding programs need to be more resistant against stress factors as heat or drought (Abberton et al., 2016).

Barley (*Hordeum vulgare* L.) and potato (*Solanum tubersosum* L.) are two of the most important crops with a world-wide production of 144 and 370 million metrics tons (FAO, 2019), respectively. Barley was firstly cultivated in the Fertile Crescent 10,000 years ago (Zohary et al., 2012). Today, it is cultivated all over the world and is mainly used for human nutrition, animal feed, and malting (Newton et al., 2011). The cultivation of barley is fundamental in the future due to its high potential to adapt to difficult conditions e.g. drought (Ceccarelli et al., 2007). In addition, it has also become an important model cereal species for research, partly because its tolerance to stress surpasses that of other major crops including wheat and rice (Nevo et al., 2012).

Potato was domesticated about 8,000 years ago in the Andes from diploid wild potatoes and became a staple food of indigenous American communities (Spooner et al., 2005). The first record of cultivated potato in Europe was on the Canary Islands in 1567 (Ríos et al., 2007) and in Spain in 1573 (Hawkes and Francisco-Ortega, 1992). Afterwards, it was adopted as a major food crop throughout Europe (Ames and Spooner, 2008). It is mainly cultivated as human food because it is important due to its high proportion of nutritional values (Jansky et al., 2019).

Breeding is responsible for 50% of crop productivity increase over the last century (Duvick, 2005; Edgerton, 2009). Therefore, breeding also needs to contribute to reach the aim of an increase of crop productivity in the future. Recent breeding applications such as genome editing (Altpeter et al., 2016) or genomic prediction (GP) (Stich and Van Inghelandt, 2018) have the potential to achieve an increase in the gain of selection. However, for both approaches, a high quality genome sequence of germplasm relevant to breeding needs to be available which is true for barley whereas for potato, a reference sequence of an European cultivar is missing.

Furthermore, in order to efficiently apply GP, a detailed understanding of genomic variation in crop genomes is fundamental. The availability of high quality genome sequences and the characterization of genomic variation can be realized due to the developments and improvements of next generation sequencing (NGS).

3.1 Improvements to sequencing technologies

In 1953, the DNA structure was discovered by Watson and Crick and from there on researchers focused on the decoding of genome sequences. Around 25 years later, the first DNA sequencing technology, also called first generation sequencing, was developed by Frederick Sanger (Sanger et al., 1977) where a read length of 1,000 bp with an accuracy of 99.99% can be obtained (Cao et al., 2017). Using this sequencing by synthesis approach, the first genomes such as that of *Escherichia coli* or yeast were sequenced (Goffeau et al., 1996; Blattner et al., 1997). Additionally, genomes of higher organisms as of *Arabidopsis thaliana* as first plant species (The Arabidopsis Genome Initiative, 2000) and the human genome (Craig Venter et al., 2001) could be sequenced using sequencing by synthesis.

In 2006, NGS, at that time developed by 454 and Illumina, also called second generation sequencing, became available and revolutionized genome research (Koboldt et al., 2013). The read length is with up to 700 bp shorter and the error rate with $\sim$1% higher than for Sanger sequencing, but the sequencing costs decreased dramatically due to the parallelism of sequencing (Tucker et al., 2009). The human genome could now be sequenced in a few days and for less than 1,000$ compared to years and billions of dollars using Sanger sequencing during the human genome project (Goodwin et al., 2016). Further advantages of NGS compared to traditional sequencing methods are e.g. sample multiplexing and higher sensitivity to detect low-frequency variants (Zhong et al., 2020).

Recently, the field of NGS was enlarged by linked-read sequencing offered by BGI (Wang et al., 2019) or formerly 10x Genomics (Weisenfeld et al., 2017). The idea behind linked-read sequencing is that paired-end short reads are derived from 50 - 100 kb DNA molecules. Ten of these molecules are partitioned into droplets and are split to smaller fragments (500 bp) where each is tagged with a 16 bp long barcode. Based on these barcoded fragments, short-read se-

quencing is performed (Elyanow et al., 2018). These short reads provide long-range information regarding to the original DNA molecule. Due to the random partition of molecules, the likelihood of assigning two molecules with the same barcode from nearby regions in the genome is very low (Elyanow et al., 2018).

Using third generation sequencing, known as long-read sequencing developed by PacBio (Wenger et al., 2019) or Oxford Nanopore (Jain et al., 2016), longer sequencing reads (10 kb - 1 Mb) can be achieved. However, this is in turn associated with high operational costs and large DNA input (PacBio) as well as high error rates (Oxford Nanopore) compared to short- and linked-read sequencing (Amarasinghe et al., 2020). This makes it less affordable for many research groups when many individuals should be sequenced. However, for the usage of single individuals, e.g. for *de novo* assemblies of new reference sequences, it is an appropriate method (Amarasinghe et al., 2020).

3.2 Crop genome assembling

In the year 2000, the first complete sequence of a plant genome, *Arabidopsis thaliana*, became available (The Arabidopsis Genome Initiative, 2000). Due to the development of NGS, the genomes of more than 100 plant species have been sequenced until the end of 2015 (Michael and VanBuren, 2015). However, many of these genome sequences are highly fragmented due to technical limitations to that time (Pham et al., 2020). Further, especially the complex nature and architecture of crop genomes characterized by polyploidy and corresponding heterozygosity (Zhang et al., 2019), a high content of transposable elements and repeats, as well as a large genome size (Mascher et al., 2017) made it difficult to assemble those genomes. However, the rapid development of new sequencing technologies such as long reads, linked reads, and proximity-ligation improved the quality of genome assembling significantly and allowed to replace established reference genome sequences by improved versions of the respective crop (Jiao and Schneeberger, 2017; Pham et al., 2020).

Despite the diploid (2n) and highly homozygous genome of barley, assembling the barley genome is a difficult challenge because of the large genome size of 5.5 Gb and the high proportion of repetitive elements (Mascher et al., 2017). However, the barley reference sequence

was first published in 2012, and was recently updated using PacBio circular consensus reads (Mascher et al., 2021). Due to the availability of this reference genome, different types of sequence variants of 23 spring barley inbreds selected from a world-wide collection (Haseneyer et al., 2010) which are the parents of a new resource for joint linkage and association mapping, the double round robin population (Casale et al., 2021), could be characterized in this thesis.
In contrast to barley, most of the available potato cultivars are tetraploid ($2n = 4x = 48$) with a high level of heterozygosity (Zhang et al., 2019). These genomic features make it difficult to assemble the genome of potato varieties. Therefore, the current potato reference sequence is that of a doubled monoploid clone from the cultivar group Phureja (Pham et al., 2020). Recently, other potato reference genomes were published, where Phureja is the pedigree (Zhou et al., 2020) or the Phureja genome sequence was used for scaffolding (van Lieshout et al., 2020). However, the cultivar group Phureja has tremendous phenotypic differences compared to the commercially established variety group Tuberosum of tetraploid cultivars (Xu et al., 2011). Additionally, preliminary comparisons between potato cultivars have shown large genomic rearrangements between them (Xu et al., 2011; Uitdewilligen et al., 2013). However, the usage of several breeding applications such as genome editing (Altpeter et al., 2016) or GP (Stich and Van Inghelandt, 2018) is facilitated by the availability of the sequence of a variety which is more related to germplasm of modern potato cultivars. Therefore, one goal of this thesis was to create a consensus reference sequence for *S. tuberosum* group Tuberosum.

3.3 Occurrence of sequence variants in crop genomes

Phenotypic variation of quantitative traits are caused by environmental and genomic factors as well as their interactions (Kearsey and Farquhar, 1998). Around the year 2000, it has been started that genetic markers such as single sequence repeats (SRR), amplified fragment length polymorphisms (AFLP), or single nucleotide variants (SNV) were examined e.g. to identify quantitative trait loci (QTL) (Kearsey and Farquhar, 1998) i.e. to identify the genomic factors underlying phenotypic variation, or to associate them with phenotypic variation (for review see Rafalski, 2010) in crop genomes. However, to that time, only a limited number, in the range of hundreds, of genetic markers could be examined (cf. Bohuon et al., 1998; Stich et al., 2006).

With the development of NGS, genomic sequencing analyses could be rendered with millions of SNV markers to study population structure and to provide a resource for phenotypic variation e.g. in *Arabidopsis thaliana* (Alonso-Blanco et al., 2016).

Beside the aforementioned genetic markers, larger genomic rearrangements have been known for a long time in the human genome, e.g. the abnormal number of chromosomes which was discovered by karyotyping (Jacobs and Strong, 1959). In crops e.g. in maize, also other microscopic genomic features such as alien chromosomes were detected by fluorescence *in situ* hybridization (Schwarzacher et al., 1992). Today, these larger genomic rearrangements, comprising of deletions, insertions, inversions, duplications, as well as translocations, are commonly defined as structural variants (SV) which are larger than 49 bp responsible for changes in the genome relative to a reference sequence or between haplotypes of a genome. These genome changes induce e.g. loss of genes, different orientation, and translocation of sequence regions (Fuentes et al., 2019). Due to those genomic characteristics, SV are supposed to play an important role contributing to phenotypic variation in crops.

Though, five years ago, genome-wide SV detection in many individuals using whole genome sequencing (WGS) was associated with high operational costs, especially for barley considering the large genome size. Thus, as an alternative to WGS and corresponding SV detection, cost-efficient mRNA sequencing could be used to examine presence/absence variation of genes what is known as dispensable transcriptome or genome (Lai et al., 2010; Hirsch et al., 2014; Jin et al., 2016). It has been accepted that a significant proportion of genes are not expressed, called expression presence/absence variation (ePAV), or are even physical absent, called genomic presence/absence variation (gPAV), in a subset of individuals of a species. In this thesis, an approach was developed where the identification of ePAV in multiple tissues could be used to determine the proportion of gPAV in the barley genome and indicated that larger regions of the genome sequence are physically absent between barley inbreds as it has been reported for maize and rice (Swanson-Wagner et al., 2006; Springer et al., 2009; Lai et al., 2010; Hirsch et al., 2014; Jin et al., 2016; Sun et al., 2018; Zhao et al., 2018).

Due to the detection of gPAV, further technical improvements, and the corresponding reduced operational costs of WGS, the genome-wide SV detection in the barley and potato genome became another objective of this thesis. Apart from the detection of presence/absence variation,

the distribution and frequency of SV in crop genomes were recently determined in rice and maize (Wang et al., 2018a; Yang et al., 2019; Kou et al., 2020). Despite the importance of barley and potato for human nutrition, the knowledge about the characterization of SV in these genomes is limited. In barley, a genome-wide study on SV was performed where the focus laid on the detection of large SV on 20 barley accessions (Jayakodi et al., 2020). In potato, SV could be extracted by the comparison of different genome assemblies (Freire et al., 2021) to detect SV between single potato clones, but, to my knowledge, only one genome-wide SV study is available where only three potato clones were examined (Lihodeevskiy and Shanina, 2021). The importance to determine the occurrence of SV in the genome has been illustrated in humans, where it has been described that SV could have an up to ∼50fold stronger influence on gene expression than SNV (Chiang et al., 2017). This is in agreement with results for different crop genomes as 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.

Detected SV in crop genomes could not only be associated with changes in gene expression but also with different phenotypic traits (for review see Saxena et al., 2014) illustrating that phenotypic variation is more likely caused by SV than by SNV which was also reported for humans (Alkan et al., 2011; Baker, 2012; Sudmant et al., 2015). Studies of discovering and association of single SV with phenotypic variation became available in crop genomes as in wheat, where single SV could be associated with traits such as flowering time (Diaz et al., 2020) or heading date (Nishida et al., 2013). Additionally, in rice, it could be associated with disease resistance and domestication (Xu et al., 2012) and in maize, it could be associated with Aluminium tolerance (Maron et al., 2013). Further, single SV could also be associated with phenotypic traits in barley such as boron toxicity tolerance (Sutton et al., 2007) as well as disease resistance (Muñoz-Amatriaín et al., 2013) and in potato, individual SV were associated with traits related to growth and development (Iovene et al., 2013). However, in these studies, only single SV were identified and associated with qualitative phenotypic traits, e.g. at the cytological level using a flurescence *in situ* hybridization based copy number variation survey (Iovene et al., 2013). Therefore, I used the SV information to examine the ability to predict quantitatively inherited phenotypic traits. Furthermore, the characterization of these SV will allow the association with

gene expression and the identification of candidate genes which underlay QTL in the barley double round robin population (Casale et al., 2021).

3.4 Benchmarking SV callers based on short-read and linked-read sequencing using computer simulations

For genome-wide SV detection, different approaches have been proposed for NGS data: genome assembling, long-read sequencing, and short-read (including linked-read) sequencing (Mahmoud et al., 2019). The genome assembly approach is a tough challenge for crop genomes, especially for barley and potato, due to the large genome size and high proportion of repetitive elements in barley (Mascher et al., 2017) or the highly heterozygous and tetraploid potato genome (Zhang et al., 2019). Additionally, due to the pairwise comparison of genome assemblies, SV between two individuals can be identified, but the time for assembling and the cost for sequencing will explode when many individuals should be assembled with a high quality. The latter is also true for long-read sequencing where the costs for sequencing are dramatically higher compared to short-read sequencing which makes it less affordable for many research groups. In contrast, short-read sequencing is a well-established tool to detect SV in human genomics (Cameron et al., 2019; Kosugi et al., 2019) and was recently considered in plant research studies (Göktay et al., 2020; Guan et al., 2021). Several signals of short-read sequencing, namely read-pair orientation, split reads, read depth, and local assembling of short reads are used to detect SV of an individual compared to a reference sequence. A described disadvantage of SV detection based on short-read sequencing is a lower mapping quality of short reads in repetitive regions (Fang et al., 2019) which were often associated with the occurrence of SV (Hu et al., 2021). This could be compensated by using linked-read sequencing where short-read sequencing is combined with long-range information. Beside using short-read sequencing signals, criteria as the density and overlap of barcodes, split molecule signals, and discrepancies in molecule coverage are used by linked-read sequencing based algorithms (Ho et al., 2020).

Due to the relatively young history of linked-read sequencing, less approaches to detect SV based on linked-read sequencing have been developed until now (for review see Ho et al., 2020),

which were mainly evaluated in the human genome. Despite the well-evaluated detection of SV based on short-read sequencing in human genomics, there are less studies available where different SV callers were evaluated in plant genomes as in rice, *Arabidopsis thaliana*, and pear (Fuentes et al., 2019; Göktay et al., 2020; Liu et al., 2020b). Additionally, it is worthwhile to evaluate SV callers with data for the specific crop, because the performance can depend on the tremendous genomic differences between crops as the genome size, repeat content, or ploidy. Therefore, in this thesis, I evaluated the performance of SV callers based on short-read and linked-read sequencing for the detection of SV in the barley and potato genome using computer simulations.

3.5 Prediction of phenotypic variation using different sequence variants as genetic markers

GP has become a powerful tool to improve the gain of selection for complex traits in animal and plant breeding programs (Meuwissen et al., 2001; Desta and Ortiz, 2014). GP works based on the usage of genetic markers to predict the breeding values of genotyped individuals. To do this, marker effects are estimated across the whole genome of those individuals based on the GP model which is trained by similar or related genotyped and phenotyped individuals of a so called training population (Desta and Ortiz, 2014). Hence, the genotyped individuals can be preselect before their phenotypes are measured in the field (Wu et al., 2022). This procedure accelerates the breeding cycle as well as reduces the cost of phenotyping (Xu et al., 2020).
In breeding programs, the classical approach is to use SNV arrays for GP (Guo et al., 2016; Crossa et al., 2017; Li et al., 2019). However, it has been reported that the genetic variance of complex traits can not be directly captured by SNV information e.g. due to high-order epistatic effects (Taylor and Ehrenreich, 2015; Wang et al., 2018b; Li et al., 2019). Thus, the prediction of complex quantitative traits could be improved by the usage of other genomic layer as larger genomic rearrangements, ePAV, or other types of sequence variants which could close the gap between genotypes and phenotypes and may even capture higher-order epistatic interactions for the prediction of phenotypic variation (Schrag et al., 2018; Hu et al., 2019; Wu et al., 2022). Thus, it is worthwhile to evaluate GP based on different sequence variants to predict different

phenotypic traits which are important for the increase of the gain of selection in the specific crop. To do this, the characterization of the barley transcriptome based on mRNA sequencing has the advantage to extract not only SNV information, but also gPAV, ePAV, and gene expression data can be examined to reduced costs compared to DNA sequencing. In contrast, DNA sequencing can be used to extract SV for GP. Therefore, in this thesis, I have evaluated the prediction of important phenotypic traits based on several sequence variants extracted from DNA and mRNA sequencing. These phenotypic traits, namely leaf angle, heading date, plant height, seed area, seed length, seed width, and thousand grain weight, are related to an increase of yield in barley.

3.6 Objectives of this thesis

The objective of my thesis was to examine the diversity and structure in the barley and potato genomes by identifying different types of sequence variants and benchmark the detection of SV in those genomes. In particular, the objectives were to

1. create a high-quality consensus reference sequence across the two haplotypes of a diploid potato clone derived from a tetraploid elite variety
2. assess sequence divergence from the available potato genome assemblies as well as among the two haplotypes
3. characterize genomic and transcriptomic variation in the barley genome using multi-tissue mRNA sequencing
4. assess the proportion of ePAV that are due to gPAV in barley
5. benchmark SV callers using simulated linked-read sequencing data in the potato genome considering different sequencing coverages, SV types, SV lengths, and haplotype incidences
6. improve SV discovery by benchmarking SV callers and their combinations with respect to their sensitivity and precision to detect SV in the barley genome
7. characterize the occurrence and distribution of SV in the genomes of 23 barley inbreds that are the parents of a resource for mapping quantitative traits, the double round robin population
8. quantify the association of SV with transcript abundance in barley
9. assess the prediction ability for quantitative phenotypic traits in barley using different sequence variants

3.7 References

Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, Costa de Oliveira A, et al. (2016), Global agricultural intensification during climate change: a role for genomics. Plant Biotechnology Journal 14:1095–1098

Alkan C, Coe BP, Eichler EE (2011), Genome structural variation discovery and genotyping. Nature Reviews Genetics 12:363–376

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, et al. (2020), Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145–161.e23

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, Cao J, et al. (2016), 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491

Altpeter F, Springer NM, Bartley LE, Blechl AE, Brutnell TP, Citovsky V, Conrad LJ, et al. (2016), Advancing crop transformation in the era of genome editing. Plant Cell 28:1510–1520

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020), Opportunities and challenges in long-read sequencing data analysis. Genome Biology 21:1–16

Ames M, Spooner DM (2008), DNA from herbarium specimens settles a controversy about origins of the European potato. American Journal of Botany 95:252–257

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

Beddington RJ, Asaduzzaman M, Clarke ME, Bremauntz AF, Guillou MD, Howlett D, Jahn MM, et al. (2012), What next for agriculture after durban? Science 335:289–290

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, et al. (1997), The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Bohuon EJ, Ramsay LD, Craft JA, Arthur AE, Marshall DF, Lydiate DJ, Kearsey MJ (1998), The association of flowering time quantitative trait loci with duplicated regions and candidate loci in *Brassica oleracea*. Genetics 150:393–401

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

Cao Y, Fanning S, Proos S, Jordan K, Srikumar S (2017), A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Frontiers in Microbiology 8:1829

Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B (2021), Genomic prediction of the recombination rate variation in barley – a route to highly recombinogenic genotypes. Plant Biotechnology Journal https://doi.org/10.1111/pbi.13746

Ceccarelli S, Grando S, Baum M (2007), Participatory plant breeding in water-limited environments. Experimental Agriculture 43:411–435

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, et al. (2017), The impact of structural variation on human gene expression. Nature Genetics 49:692–699

Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. (2001), The sequence of the human genome. Science 291:1304–1351

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, et al. (2017), Genomic selection in plant breeding: Methods, models, and perspectives. Trends in Plant Science 22:961–975

Desta ZA, Ortiz R (2014), Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science 19:592–601

Diaz S, Ariza-Suarez D, Izquierdo P, Lobaton JD, De La Hoz JF, Acevedo F, Duitama J, et al. (2020), Genetic mapping for agronomic traits in a MAGIC population of common bean (*Phaseolus vulgaris* L.) under drought conditions. BMC Genomics 21:799

Duvick DN (2005), The contribution of breeding to yield advances in maize (*Zea mays* L.). Advances in Agronomy 86:83–145

Edgerton MD (2009), Increasing crop productivity to meet global needs for feed, food, and fuel. Plant Physiology 149:7–13

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

Fang L, Kao C, Gonzalez MV, Mafra FA, Pellegrino da Silva R, Li M, Wenzel SS, et al. (2019), LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. Nature Communications 10:5585

Freire R, Weisweiler M, Guerreiro R, Baig N, Hüttel B, Obeng-Hinneh E, Renner J, et al. (2021), Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. G3 Genes|Genomes|Genetics 11:jkab330

Frona D, Janos S, Harangi-Rakos M (2019), The challenge of feeding the world. Sustainability 11:5816

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, et al. (2019), Structural variants in 3000 rice genomes. Genome Research 29:870–880

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, et al. (1996), Life with 6000 genes. Science 274:546–567

Göktay M, Fulgione A, Hancock AM (2020), A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. Molecular Biology and Evolution 38:1498–1511

Goodwin S, McPherson JD, McCombie WR (2016), Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics 17:333–351

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

Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D (2016), Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theoretical and Applied Genetics 129:2413–2427

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 Breeding 129:271–279

Hawkes JG, Francisco-Ortega J (1992), The potato in Spain during the late 16th century. Economic Botany 46:86–97

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, et al. (2014), Insights into the maize pan-genome and pan-transcriptome. The Plant Cell 26:121–135

Ho SS, Urban AE, Mills RE (2020), Structural variation in the sequencing era. Nature Reviews Genetics 21:171–189

Hu X, Xie W, Wu C, Xu S (2019), A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnology Journal 17:2011–2020

Hu Y, Colantonio V, Müller BS, Leach KA, Nanni A, Finegan C, Wang B, et al. (2021), Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nature Communications 12:1227

Iovene M, Zhang T, Lou Q, Buell CR, Jiang J (2013), Copy number variation in potato - an asexually propagated autotetraploid species. Plant Journal 75:80–89

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

Jain M, Olsen HE, Paten B, Akeson M (2016), The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 17:239

Jansky S, Navarre R, Bamberg J (2019), Introduction to the special issue on the nutritional value of potato. American Journal of Potato Research 96:95–97

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, et al. (2020), The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289

Jiao WB, Schneeberger K (2017), The impact of third generation genomic technologies on plant genome assembly. Current Opinion in Plant Biology 36:64–70

Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y (2016), Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Nature Scientific Reports 6:18936

Kearsey MJ, Farquhar AG (1998), QTL analysis in plants; where are we now? Heredity 80:137–142

Koboldt D, Steinberg K, Larson D, Wilson R, Mardis ER (2013), The next-generation sequencing revolution and its impact on genomics. Cell 1551:27–38

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 Biology 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. Molecular Biology and Evolution 37:3507–3524

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, et al. (2010), Genome-wide patterns of genetic variation among elite maize inbred lines. Nature Genetics 42:1027–1030

Li Z, Gao N, Martini JW, Simianer H (2019), Integrating gene expression data into genomic prediction. Frontiers in Genetics 10:126

van Lieshout N, van der Burgt A, de Vries ME, ter Maat M, Eickholt D, Esselink D, et al. (2020), Solyntus, the new highly contiguous reference genome for potato (*Solanum tuberosum*). G3 Genes|Genomes|Genetics 10:3489–3495

Lihodeevskiy GA, Shanina EP (2021), Structural variations in the genome of potato varieties of the ural selection. Agronomy 11:1703

Liu M, Li Y, Ma Y, Zhao Q, Stiller J, Feng Q, Tian Q, Liu D, Han B, Liu C (2020a), The draft genome of a wild barley genotype reveals its enrichment in genes related to biotic and abiotic stresses compared to cultivated barley. Plant Biotechnology Journal 18:443–456

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

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 Biology 20:246

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, et al. (2013), Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proceedings of the National Academy of Sciences of the United States of America 110:5241–5246

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, et al. (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, et al. (2021), Long-read sequence assembly: a technical evaluation in barley. The Plant Cell 33:1888–1906

Meuwissen TH, Hayes BJ, Goddard ME (2001), Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Michael TP, VanBuren R (2015), Progress, challenges and the future of crop genomes. Current Opinion in Plant Biology 24:71–81

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, et al. (2013), Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biology 14:R58

Nevo E, Fu YB, Pavlicek T, Khalifa S, Tavasi M, Beiles A (2012), Evolution of wild cereals during 28 years of global warming in Israel. Proceedings of the National Academy of Sciences 109:3412–3415

Newton AC, Flavell AJ, George TS, Leat P, Mullholland B, Ramsay L, Revoredo-Giha C, et al. (2011), Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. Food Security 3:141–178

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. Molecular Breeding 31:27–37

Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, Ou S, et al. (2020), Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience 9:1–11

Rafalski JA (2010), Association genetics in crop improvement. Current Opinion in Plant Biology 13:174–180

Ríos D, Ghislain M, Rodríguez F, Spooner DM (2007), What is the origin of the European potato? Evidence from Canary Island landraces. Crop Science 47:1271–1280

Sanger F, Nicklen S, Coulson AR (1977), DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74:5463–5467

Saxena RK, Edwards D, Varshney RK (2014), Structural variations in plant genomes. Briefings in Functional Genomics and Proteomics 13:296–307

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

Schwarzacher T, Anamthawat-Jónsson K, Harrison GE, Islam AK, Jia JZ, King IP, Leitch AR, et al. (1992), Genomic in situ hybridization to identify alien chromosomes and chromosome segments in wheat. Theoretical and Applied Genetics 84:778–786

Spooner DM, McLean K, Ramsay G, Waugh R, Bryan GJ (2005), A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proceedings of the National Academy of Sciences of the United States of America 102:14694–14699

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, et al. (2009), Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLOS Genetics 5:1–17

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, Van Der Voort JR, Peleman J, Sørensen AP, Reif JC (2006), Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Molecular Breeding 17:217–226

Stich B, Van Inghelandt D (2018), Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. Frontiers in Plant Science 9:159

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. (2015), An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, et al. (2018), Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nature Genetics 50:1289–1295

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, et al. (2007), Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318:1446–1449

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS (2006), All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proceedings of the National Academy of Sciences 103:6805–6810

Taylor MB, Ehrenreich IM (2015), Higher-order genetic interactions and their contribution to complex traits. Trends in Genetics 31:34–40

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

The Food and Agriculture Organization (FAO) (2019), http://www.fao.org/faostat/en/#data/QV

Tucker T, Marra M, Friedman JM (2009), Massively parallel sequencing: the next big thing in genetic medicine. American Journal of Human Genetics 85:142–154

Uitdewilligen JG, Wolters AMA, D'hoop BB, Borm TJ, Visser RG, van Eck HJ (2013), A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLOS ONE 8:e62355

Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, et al. (2019), Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nature Genetics 51:224–229

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, et al. (2018a), Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 5577703:43–49

Wang X, Xu Y, Hu Z, Xu C (2018b), Genomic selection methods for crop improvement: current status and prospects. Crop Journal 6:330–340

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

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. (2019), Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37:1155–1162

Wu PY, Stich B, Weisweiler M, Shrestha A, Erban A, Westhoff P, Van Inghelandt D (2022), Improvement of prediction ability by integrating multi-omic datasets in barley. BMC Genomics 23:200

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, et al. (2012), Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology 30:105–111

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

Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, et al. (2020), Enhancing genetic gain through genomic selection: from livestock to plants. Plant Communications 1:100005

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, et al. (2019), Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nature Genetics 51:1052–1059

Zhang C, Wang P, Tang D, Yang Z, Lu F, Qi J, Tawari NR, et al. (2019), The genetic basis of inbreeding depression in potato. Nature Genetics 51:374–378

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, et al. (2015), Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27:1595–1604

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, et al. (2018), Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature Genetics 50:278–284

Zhong Y, Xu F, Wu J, Schubert J, Li MM (2020), Application of next generation sequencing in laboratory medicine. Annals of Laboratory Medicine 41:25–43

Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, et al. (2020), A platinum standard pan-genome resource that represents the population structure of Asian rice. Scientific Data 7:113

Zohary D, Hopf M, Weiss E, Zohary D, Weiss E, Hopf M (2012), Domestication of plants in the old world - the origin and spread of domesticated plants in south-west Asia, Europe, and the Mediterranean Basin

4. Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety

This manuscript was published in G3 Genes|Genomes|Genetics in September, 2021.

Authors:

Ruth Freire, Marius Weisweiler, Ricardo Guerreiro, Nadia Baig, Bruno Hüttel, Evelyn Obeng-Hinneh, Juliane Renner, Stefanie Hartje, Katja Muders, Bernd Truberg, Arne Rosen, Vanessa Prigge, Julien Bruckmüller, Jens Lübeck, Benjamin Stich.

Contribution: Shared first author

Benjamin Stich designed and coordinated the project.

Marius Weisweiler, Ruth Freire, Ricardo Guerreiro, and Nadia Baig performed the analyses.

Marius Weisweiler, Ruth Freire, Ricardo Guerreiro, and Benjamin Stich wrote the manuscript.

Evelyn Obeng-Hinneh, Juliane Renner, Stefanie Hartje, Katja Muders, Bernd Truberg, Arne Rosen, Vanessa Prigge, Julien Bruckmüller, and Jens Lübeck provided the genetic material.

Bruno Hüttel extracted DNA and prepared the libraries.

G3, 2021, **11(12),** jkab330
DOI: 10.1093/g3journal/jkab330
Advance Access Publication Date: 17 September 2021
Genome Report

Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety

Ruth Freire,[1,†] Marius Weisweiler ,[1,†] Ricardo Guerreiro,[1,†] Nadia Baig,[1] Bruno Hüttel,[2] Evelyn Obeng-Hinneh,[3] Juliane Renner,[3] Stefanie Hartje,[3] Katja Muders,[4] Bernd Truberg,[4] Arne Rosen,[4] Vanessa Prigge,[5] Julien Bruckmüller ,[6] Jens Lübeck,[6] and Benjamin Stich[1,7,*]

[1]Institute for Quantitative Genetics and Genomics of Plants, 40225 Düsseldorf, Germany
[2]Max Planck-Genome-Centre Cologne, Max Planck Institute for Plant Breeding, 50829 Köln, Germany
[3]Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, 17111 Hohenmocker, Germany
[4]Nordring-Kartoffelzucht- und Vermehrungs-GmbH, 18190 Sanitz, Germany
[5]SaKa Pflanzenzucht GmbH & Co. KG, 24340 Windeby, Germany
[6]Solana Research GmbH, 24340 Windeby, Germany and
[7]Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, 40225 Düsseldorf, Germany
[†]These authors contributed equally to this work.

*Corresponding author: Email: benjamin.stich@hhu.de

Abstract

Potato (*Solanum tuberosum* L.) is one of the most important crops with a worldwide production of 370 million metric tons. The objectives of this study were (1) to create a high-quality consensus sequence across the two haplotypes of a diploid clone derived from a tetraploid elite variety and assess the sequence divergence from the available potato genome assemblies, as well as among the two haplotypes; (2) to evaluate the new assembly's usefulness for various genomic methods; and (3) to assess the performance of phasing in diploid and tetraploid clones, using linked-read sequencing technology. We used PacBio long reads coupled with 10x Genomics reads and proximity ligation scaffolding to create the dAg1_v1.0 reference genome sequence. With a final assembly size of 812 Mb, where 750 Mb are anchored to 12 chromosomes, our assembly is larger than other available potato reference sequences and high proportions of properly paired reads were observed for clones unrelated by pedigree to dAg1. Comparisons of the new dAg1_v1.0 sequence to other potato genome sequences point out the high divergence between the different potato varieties and illustrate the potential of using dAg1_v1.0 sequence in breeding applications.

Keywords: reference sequence; elite potato variety; chromosome-scale; genome divergence; intragenomic diversity

Introduction

Potato (*Solanum tuberosum* L.) was domesticated about 8000 years ago in the Andes from diploid wild potatoes and became a staple food of indigenous American communities (Spooner *et al.* 2005). Because of its high nutritional value (Jansky *et al.* 2019), the potato is nowadays one of the most important crops for humanity and its global production exceeds 370 million metric tons (FAO 2019).

The number of potato cultivars is in the thousands (FAO 2008), most of which are tetraploid ($2n = 4\times = 48$), with a high level of heterozygosity and strong inbreeding depression (Zhang *et al.* 2019). With the steady rise of the human population, and growing fears of food insecurity (Beddington 2010), it is crucial to increase potato productivity. Inter alia, considerable increases are expected to be contributed by plant breeding (Lenaerts *et al.* 2019). Modern breeding tools such as genome editing (Altpeter *et al.* 2016) and genomic selection (Stich and Van Inghelandt 2018) have the potential to enhance the gain of selection in potato. However, to utilize the full potential of these tools, high-quality reference genomes of germplasm relevant to breeding are required.

The current *S. tuberosum* reference genome is that of a doubled monoploid clone from the cultivar group Phureja (Xu *et al.* 2011; Sharma *et al.* 2013; Pham *et al.* 2020). However, group Phureja has considerable genome and phenotype differences compared to the commercially established group Tuberosum of tetraploid cultivars (Xu *et al.* 2011), which makes it presumably not ideal as a reference for the latter. Furthermore, preliminary comparisons between cultivars indicated substantial sequence and structural variations (SV; Xu *et al.* 2011; Uitdewilligen *et al.* 2013), which calls for cultivar-specific genome assemblies as to optimally exploit genomic tools for potato breeding.

Assembling potato genomes is challenging because of their high levels of heterozygosity. Mixed heterozygous and homozygous regions make it difficult for algorithms to find a single unique path of overlapping reads, leading to more fragmented assemblies and a requirement of higher sequencing coverage (Pryszcz and Gabaldón 2016). If heterozygosity is very high, the alternative haplotype contigs are assumed to be separate regions of the genome, a phenomenon called undercollapsed heterozygosity (Matthews *et al.* 2018). This effect is more pronounced in

Received: May 31, 2021. **Accepted:** September 08, 2021

tetraploid genomes, since they have more alternative haplotype versions of the same region. Long-read sequencing technologies such as PacBio (Shearman *et al.* 2020; Vollger *et al.* 2020) and Nanopore (Kuderna *et al.* 2019; Low *et al.* 2019; Kinkar *et al.* 2021) aim to overcome the problem of heterozygosity, allowing more space for overlaps during assembly (Jiao and Schneeberger 2017). Further sequencing technologies such as proximity ligation and optical mapping help resolving areas that are difficult to assemble (Field *et al.* 2020).

In recent years, potato genome assemblies of wild diploid potato relatives Solanum *commersonii* (Aversano *et al.* 2015) and Solanum *chacoense* (M6; Leisner *et al.* 2018) have become available. Exploiting the latest sequencing technological advances, Zhou *et al.* (2020) assembled the phased genome sequence of RH89-039-16 (RH89), a diploid clone derived from a cross between *S. tuberosum* dihaploid and a diploid clone, which in turn was generated from a cross between two *S. tuberosum* group Phureja hybrids (Xu *et al.* 2011). Finally, the first non-Phureja *S. tuberosum* assembly has been recently published (Solyntus_v1.1; van Lieshout *et al.* 2020). For the latter, however, the Phureja genome (DM_v4.03) has been used for reference-based scaffolding. Therefore, to our knowledge, no genome sequence of an elite variety is available nor any pure chromosome-level assembly of *S. tuberosum* group Tuberosum.

The objectives of this study were (1) to create a high-quality consensus sequence across the two haplotypes of a diploid clone derived from a tetraploid elite variety and assess the sequence divergence from the available potato genome assemblies as well as among the two haplotypes; (2) to evaluate the new assembly's usefulness for various genomic methods; and (3) to assess the performance of phasing in diploid and tetraploid clones using linked-read sequencing technology.

Materials and methods

Genetic material, DNA, and RNA extraction

Three gynogenic dihaploid *S. tuberosum* clones (dAg1, dAg2, and dAg3) were created from *S. tuberosum* group Tuberosum tetraploid cv. Agria (tAg). The haploid inducer was *S. tuberosum* group Phureja IVP06-153. Besides tAg, its parental clones tPa1 and tPa2 as well as five tetraploid elite potato clones (tV1–tV5) were included in this study. DNA was extracted from the leaves of all clones according to Mayjonade *et al.* (2016). For RNA sequencing, 10 tubers of tAg were grown in a cultivation chamber set to 25°C during day (6–22 h) and 20°C during night. The light intensity was about 300 µmol/m^{2}s in the leaf canopy. Samples of leaves, stolons, and flowers were harvested at 15 (leaves and stolons) and 45 (flowers) days after planting. Total RNA was extracted using RNeasy Plant MiniKit (Qiagen, Hilden, Germany) following the manufacturer's instructions. RNA was pooled to equal concentration for the following library preparation.

Preparation of libraries and sequencing

For all clones, 10x Genomics (10xG; Pleasanton, CA, USA) libraries were prepared (Supplementary Table S1) following the manufacturer's recommendations, using 1 ng of DNA input, where size selection was performed before library preparation on BluePippin (SAGE Sciences, Beverly, MA, USA) with a high-pass protocol allowing a size selection start at 40 kb. The quantity and quality control of size-selected DNA were performed with Qubit (Thermo) and with a Genomic tape (Agilent TapeStation). Sequencing of 10xG libraries was performed on an Illumina (San Diego, CA, USA) HiSeq3000 in paired-end read mode.

For dAg1, SMRTbell libraries were prepared as recommended by Pacific Biosciences (Menlo Park, CA, USA, SMRTbell Template Prep Kit 1.0-SPv3), including a final size selection on Blue Pippin to remove fragments lower than 10 kb. Sequencing was performed on a PacBio Sequel I with Binding Kit 2.0 and Sequencing chemistry 2.0 for 10 h or Binding Kit 3.0 and Sequencing chemistry 3.0 for 20 h, as recommended by Pacific Biosciences.

Proximity ligation (Hi-C) data were generated for dAg1 by Dovetail (Boston, MA, USA), following the protocol of Lieberman-Aiden *et al.* (2009). A total of 129 × 10^6 2 × 150 bp Hi-C reads were sequenced.

Pooled RNA from leaves, stolons, and flowers was used to prepare an Iso-Seq library following manufacturer instructions. Sequencing was performed on the PacBio Sequel II using the Sequel II Sequencing Kit 2.0 chemistry. Iso-seq v3 pipeline (https://github.com/PacificBiosciences/IsoSeq) was used to generate final RNA sequencing data.

Genome assembly

Our objective was to create one contiguous consensus assembly across the two haplotypes of dAg1 and phase the existing intragenomic variants for diploid and tetraploid clones in a second step. We have evaluated two different assembly strategies to obtain the dAg1_v1.0 genome sequence (Supplementary Figure S1), but in this manuscript, only the final assembly strategy and results are presented.

Final assembly strategy: PacBio assembly as backbone

All PacBio reads that had ≤200× coverage, an error rate <15% after error correction, and a length ≥1000 bp were assembled with Canu v1.8 (Koren *et al.* 2017). Parallelly, the same reads were also assembled using Falcon and Falcon-unzip (Chin *et al.* 2016). To deal with the higher error rate of PacBio reads, both assemblies were polished using Pilon (v1.22; Walker *et al.* 2014) with the less error-prone 10xG linked reads, where mapping was performed with longranger align (v2.2.2) (Zheng *et al.* 2016). Furthermore, the polished Canu assembly was filtered with Purge Haplotigs (Roach *et al.* 2018) to avoid undercollapsed heterozygosity (Matthews *et al.* 2018) by discarding alternative haplotigs.

A hybrid assembly was created using quickmerge (v0.3; Chakraborty *et al.* 2016), where the polished Falcon assembly was used as reference and the polished and deduplicated Canu assembly as query. This was followed by a second round of Pilon polishing with mapped 10xG linked reads. These mapped reads were additionally used to correct misassemblies using Tigmint (Jackman *et al.* 2018) and the assembly was filtered with Purge Haplotigs. Arcs (v1.0.6; Yeo *et al.* 2018) and Links (v1.8.7; Warren *et al.* 2015) were used to scaffold contigs of the polished, corrected, deduplicated quickmerge assembly with the 10xG library 1, lowering the minimum aligned reads to 3 instead of 5 (-c 3) and using k-mers of size 20 (-k 20). Thereafter, the step was iterated with 10xG library 2. Finally, a last round of polishing with Pilon and filtering with Purge Haplotigs was performed.

Hi-C scaffolding

The reads of the Hi-C library were mapped against the scaffolded hybrid assembly in two steps with different software. In the first step, we used BWA-MEM (v0.7.15; Li and Durbin 2010) for mapping and Salsa (Ghurye *et al.* 2017) for scaffolding, with misassembly correction activated (-m yes). In the second step, Juicer was used for mapping (Durand *et al.* 2016) and 3D-DNA (Dudchenko *et al.* 2017, 2018) for scaffolding. Contigs smaller

than 12.5 kb were ignored during scaffolding and the repeat coverage misjoin threshold (–editor-repeat-coverage) was set to 3. The resulting contact maps were visualized using Juicebox (Durand *et al.* 2016; Dudchenko *et al.* 2017, 2018) and a final manual curation and scaffolding were performed.

Evaluation of assemblies

A custom python script was used at all steps of the assembly to obtain several statistics, namely the N50, N90, L50, L90, number of Ns per 100 kb, as well as scaffold number, and total sequence length. Benchmarking Universal Single-Copy Orthologs (BUSCO; Simão *et al.* 2015) were used to assess gene completeness compared to the Solanaceae gene set (odb10; Kriventseva *et al.* 2019).

Whole-genome alignments of the final dAg1_v1.0 assembly and the four assemblies DM_v4.04, DM_v6.1, RH89, and Solyntus_v1.1 were performed with nucmer (-l 1000 -c 1000 -d 10) from the MUMMER package (v4.0.0beta2; Marçais *et al.* 2018). Additionally, for dAg1_v1.0 *vs* Solyntus_v1.1, a second alignment was performed using a lower minimum length of single exact matches (-l 100) and of a cluster of matches (-c 100) to visualize the alignment.

In order to evaluate our final assembly, mapping of 10xG linked reads from various diploid and tetraploid clones against our and the existing potato reference assemblies (dAg1_v1.0, DM_v4.04, DM_v6.1, RH89, M6, and Solyntus_v1.1) was performed using longranger align. Illumina sequencing data of the diploid wild potato species (*Solanum bukasovii*, dW; Kyriakidou *et al.* 2020) were downloaded from the SRA database and mapped against the six genomes using BWA-MEM. Samtools (v1.10; Li *et al.* 2009) was used to calculate the proportion of mapped reads and properly paired reads (Thankaswamy-Kosalai *et al.* 2017).

Gene annotation

The MAKER pipeline (Campbell *et al.* 2014) was used to annotate genes. A custom repeat library was created with Repeatmodeler (Smit *et al.* 2013) and Mite Hunter (Han and Wessler 2010) according to Campbell *et al.* (2014). Repeatmasker (Smit *et al.* 2013) was then used to mask these repeat regions in the genome. The Iso-seq RNA data generated for tAg in this project as well as published mRNA reads (SRX4882701; Caruana *et al.* 2019) from tAg, assembled into a transcriptome with Trinity (v2.11.0) (Haas *et al.* 2013), were used as EST evidences. Protein evidences were UniProt proteins of Solanum (The UniProt Consortium 2019). Snap (Korf 2004) and Augustus (Stanke *et al.* 2008) were used as gene predictors. Orthologous analysis with UniProt proteins of Solanum was done with Orthofinder (Emms and Kelly 2019).

Iso-seq RNA analysis

High-quality RNA reads obtained with Iso-seq version 3 pipeline (https://github.com/PacificBiosciences/IsoSeq) for tAg were mapped against dAg1_v1.0, DM_v4.04, DM_v6.1, RH89, M6, and Solyntus_v1.1 genomes using Minimap2 (Li 2018). Mapped reads against dAg1_v1.0, DM_v6.1, and RH89 were then filtered for alignments with ≥99% coverage and ≥95% identity. Redundant isoforms were removed using cDNA-Cupcake pipeline (http://github.com/Magdoll/cDNA_Cupcake). Collapsed isoforms were categorized according to dAg1_v1.0, DM_v6.1, and RH89 annotations by using SQANTI3 (Tardaguila *et al.* 2018). Alternative splicing was investigated with SUPPA2 (Trincado *et al.* 2018).

Variant calling, phasing, and annotation

The dAg1_v1.0 assembly was used as reference to call single nucleotide variants (SNV), and small insertions and deletions (indels, <50 bp) for all potato clones. The corresponding 10xG linked reads of the diploid clones dAg1, dAg2, and dAg3 were aligned with longranger wgs (v2.2.2), and phased SNV and indels were called using freebayes (v1.3.2-40; Garrison and Marth 2012). 10xG linked reads of the three tetraploid clones tAg, tPa1, and tPa2 were mapped against the dAg1_v1.0 assembly using longranger align (v2.2.2) and variants were called by freebayes. Variants of the samples dAg1, dAg2, dAg3, tAg, tPa1, and tPa2 were filtered for a minimum depth of 10. The variants of the clones were phased with whatshap polyphase (Schrinner *et al.* 2020). The allele profiles of regions for which phase information was available for the offspring (dAg1, dAg2, dAg3, and tAg) were compared with that of the respective parents (tAg, tPa1, and tPa2). The proportion of regions with correctly phased allele profiles in the offspring compared to the allele profiles of the parental clones was calculated.

Sorting Intolerant From Tolerant 4G (SIFT4G, v2.4) was used to annotate tolerant (score >0.05) and deleterious (score ≤0.05) variants based on the conversion of amino acid sequences (Vaser *et al.* 2016). The SIFT4G database was built using SIFT4_Create_Genomic_DB with the uniref90 database, the dAg1_v1.0 sequence, and its corresponding predicted genes and proteins. The number of genes with at least one putative deleterious variant was estimated.

Pericentromeric regions of the potato chromosomes of DM_v4.03 were determined based on the recombination rates reported for the DRH population (Manrique-Carpintero *et al.* 2016). Thereafter, we determined the pericentromeric regions in the dAg1_v1.0 sequence based on the coordinates of the whole-genome alignment between dAg1_v1.0 and DM_v4.03 using show-coords from the MUMMER package. We then used a *t*-test to examine the difference of the proportion of genes with at least one deleterious variant between ∅dAg1-3 and ∅Pa1-2 in pericentromeric to subtelomeric regions for its statistical significance. Additionally, a *t*-test was used to test for a mean difference of the proportion of genes with at least one deleterious variant, calculated in 1-Mb windows across the genome, between diploid (∅dAg1-3) and tetraploid clones (∅tPa1-2).

SV between the two haplotypes of dAg1 were identified from PacBio reads, using the CuteSV algorithm (v1.0.8; Jiang *et al.* 2020) after mapping the reads with Minimap2. Sequence divergence between the assembly sequences was estimated as the proportion of the number of bp affected by SV, where the latter was extracted from the whole-genome alignments (-l 1000 -c 1000 -d 10), including all final primary scaffolds of the four potato genomes using show-diff from the MUMMER package.

Results and discussion

Genome assembly

Two PacBio assemblies were created for the final assembly strategy: the first with Canu comprising 14,037 contigs and the second with Falcon comprising 2,109 contigs. The Canu assembly had a larger than expected assembly size and the BUSCO analysis indicated a high proportion of duplications, both signs of undercollapsed heterozygosity. In addition, the N50 value of the Falcon assembly (0.618 Mb) was higher than that of the Canu assembly (0.203 Mb). Consequently, the Falcon assembly was used as reference and the Canu assembly as query in creating the hybrid assembly. The resulting hybrid assembly had a reduced number of contigs (1,592) and the N50 increased to 0.865 Mb. After two rounds of 10xG scaffolding, the number of scaffolds decreased to 704 and the N50 increased to 1.656 Mb, where, for the Solanaceae gene set, a BUSCO statistic of 95% was observed.

Table 1 Assembly statistics of different steps of our final genome assembly strategy for dAg1

Assembly step	No. of contigs	Assembly size (Mb)	Largest contig (Mb)	N50 (Mb)	N90 (Mb)	L50	L90	Ns per 100kb	BUSCO (%)
Assembling									
Canu	14,037	1,343.9	4.559	0.203	0.035	1,643	7,787	0	95
Falcon	2,109	845.7	4.904	0.618	0.206	393	1,315	0	95
quickmerge	1,592	889.7	10.609	0.865	0.276	267	974	0	95
Arcs 1×	1,055	895.9	13.589	1.440	0.407	176	635	757	95
Arcs 2×	704	788.1	13.585	1.656	0.548	136	445	977	95
Hi-C scaffolding									
Hi-C Salsa	385	788.4	29.219	5.059	1.007	41	175	1,006	95
Hi-C 3D-DNA	12 (+614)	812.2	89.719	57.412	52.458	6	12	994	94

For details see *Materials and Methods*.

This assembly strategy based on PacBio long reads as backbone resulted in a more contiguous assembly with a closer-to-expected assembly size and a lower number of Ns (Table 1) compared to another strategy with 10xG linked reads as backbone (Supplementary Text and Table S2). Therefore, this former assembly was used for further scaffolding using Hi-C data. The two Hi-C scaffolding approaches led to drastically increased N50 values up to 57.4 Mb and a slight decrease in BUSCO statistics. The latter phenomenon was already observed by Kadota *et al.* (2020) and might be due to misassembly over-correction and/or imperfect manual curation at the last stage.

The unscaffolded minor contigs were concatenated as ChrUn (65.88 Mb), which represents 8.3% of the genome and hosts 2,124 (4.7%) of the annotated genes. When ignoring Ns and including ChrUn, the final assembly size of the dAg1_v1.0 genome was 812 Mb, which is in a similar range compared to 731, 807, and 1,674 Mb (diploid genome size) for DM_v6.1, M6, and RH89, respectively (Supplementary Table S3). Considering only the 12 final scaffolds as chromosomes, the genome size of 744 Mb is larger than in M6, Solyntus_v1.1, and DM_v6.1 (499, 716, and 731 Mb).

Evaluation of dAg1_v1.0 genome sequence

The visual inspection of the Hi-C contact maps of the dAg1_v1.0 sequence suggested the presence of clear contact areas between the ends of all chromosomes (Figure 1A). This has been observed earlier for plant genome (Liu and Weigel 2015; Liu *et al.* 2017) and supports the quality of our assembly. As additional quality control, especially to evaluate the successful purging of the second haplotype from the consensus assembly, we have evaluated the genome-wide distribution of the read depth. Only a few coverage spikes were observed for the final assembly (Supplementary Figure S2). A first analysis of these regions with particularly high coverage suggests that they are related to repetitive sequences. These attributes indicate that our assembly has a high quality.

Visual inspections of the dot plots of whole-genome alignments between dAg1_v1.0 *vs* DM_v6.1 and between dAg1_v1.0 *vs* RH89 (Figure 1, B and C) suggested a high level of correspondence between dAg1_v1.0 and the other two potato genomes. A reduction of the minimum length of a single exact alignment match from 1,000 to 100 bp was necessary to visualize the whole-genome alignment of dAg1_v1.0 and Solyntus_v1.1 (Figure 1D) which suggests a lower level of correspondence, which might be explained by misassemblies in the Solyntus_v1.1 genome scaffolded by DM_v4.03 (DM_v4.04 without unscaffolded contigs). This previous Phureja genome presumably included misassemblies and sequencing errors due to the limited sequencing resources available in 2011 when the genome was assembled (Pham *et al.* 2020). This explanation is supported by the observation of similar differences in the whole-genome alignment between dAg1_v1.0 and DM_v4.04 (Supplementary Figure S3) but not between dAg1_v1.0 and DM_v6.1.

The reason for larger gaps in the abovementioned whole-genome alignments might be a lower assembly quality in these regions, especially in the pericentromeric regions. The latter regions are characterized by high repeat frequencies which are difficult to assemble.

Visual inspections of the dot plots of whole-genome alignments of RH89 *vs* dAg1_v1.0 and of RH89 *vs* DM_v6.1 (Supplementary Figure S4) suggested that the RH89 and DM_v6.1 genomes are more similar than the RH89 and dAg1_v1.0 genomes. This visual impression is supported by the observation of a lower sequence divergence of ~8.5% between RH89 and DM_v6.1 compared to ~10.8% between RH89 and dAg_v1.0, calculated based on the sequence differences due to SV. This finding is in agreement with the RH89 pedigree, which implies a higher relatedness between RH89 and DM than between RH89 and dAg. The sequence divergence between dAg_v1.0 and DM_v6.1 and Solyntus_v1.1 was 8.2% and 8.6%, respectively.

To assess the completeness and correctness of the dAg1_v1.0 sequence relative to other potato sequences, we mapped 10xG linked reads of various potato varieties (tV1–tV5) and one wild species (dW) to dAg1_v1.0, DM_v4.04, DM_v6.1, RH89, Solyntus_v1.1, and M6 genome sequences. The percentage of mapped reads of all examined clones was higher against *S. tuberosum* reference sequences than against M6 (Figure 2A). The proportion of mapped reads against dAg1_v1.0 was high and similar to the other *S. tuberosum* genomes. However, the proportion of properly paired reads, considered to be a more accurate quality measure (Thankaswamy-Kosalai *et al.* 2017), was on average across all examined clones the highest for the dAg1_v1.0 genome (Figure 2B). PacBio and Iso-seq reads from diploid and tetraploid Agria were also mapped against the reference assemblies and high percentages of reads were mapped in all cases (Figure 3). These observations together indicated the high completeness and especially the correctness of the dAg1_v1.0 genome assembly, which will therefore be highly useful for genome-assisted breeding applications in potato and the basis for many future research projects on diploid and tetraploid potato.

Transcript analysis

To illustrate further the usefulness of the dAg1_v1.0 genome for research on tetraploid potatoes, a transcript analysis was performed. High-quality Iso-seq reads of tAg were mapped

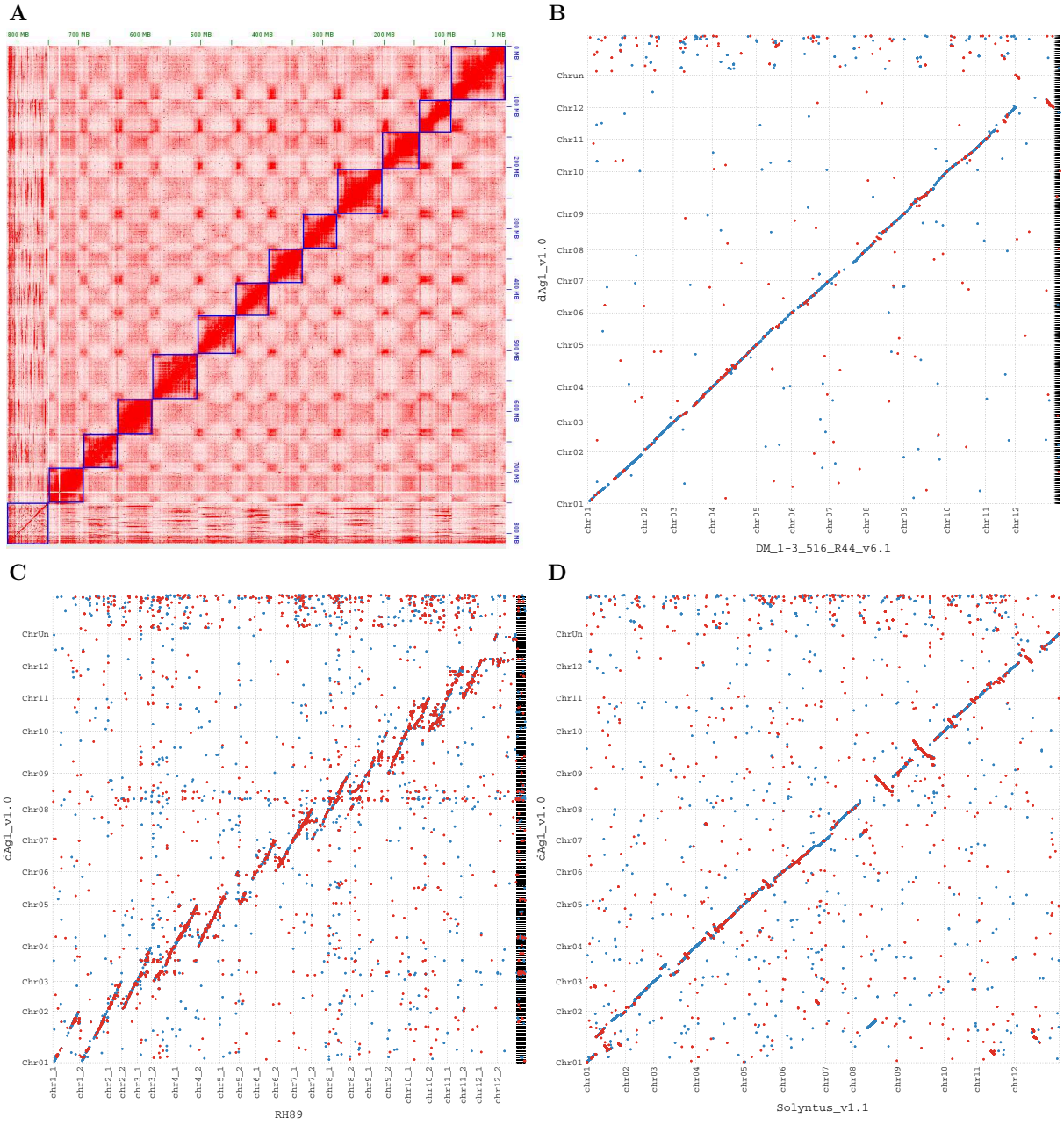

Figure 1 Hi-C contact map of dAg1_v1.0 sequence (A). Dot plots of whole-genome alignments of dAg1_v1.0 (vertical) *vs* DM_v6.1 (B), RH89 (C), and Solyntus_v1.1 (D) genomes (horizontal). Each dot indicates an alignment with a length of ≥1000 bp between the two genomes (≥100 bp for D). Forward and reverse alignments are represented as blue and red dots, respectively.

against dAg1_v1.0, DM_v6.1, and RH89 genomes. After collapsing the mapped reads, unique transcripts were compared with the dAg1_v1.0 annotation set as well as those published for DM_v6.1 and RH89 (Pham *et al.* 2020; Zhou *et al.* 2020). SQANTI3 analyses showed a higher ratio of transcripts associated with annotated genes compared to novel genes in dAg1_v1.0 and a higher number of annotated genes in dAg1_v1.0 compared to the other genomes (Supplementary Figure S5A). Though this finding may be biased by the fact that Iso-seq reads were used as evidences in obtaining gene models in dAg1_v1.0, the results of SQANTI3 suggest a good quality of annotation for dAg1_v1.0.

Iso-seq reads were obtained from an RNA bulk of leaves, stolons, and flowers. So a broad representation of genes should be expected. Nevertheless, some tissues, like tubers, are not present in Iso-seq reads. To assess whether dAg_v1.0 annotation presents some bias, an orthologous analysis was done between dAg_v1.0 genes and Solanum proteins from the UniProt database. Up to 90% of the tuber proteins have an orthologous in dAg_v1.0 genes, the same proportion found in RH89 and DM_v6.1. So, despite the absence of tuber Iso-seq reads, no clear bias was found in dAg_v1.0 annotated genes.

The proportion of genes with more than one isoform was greater than the number of genes with only one isoform

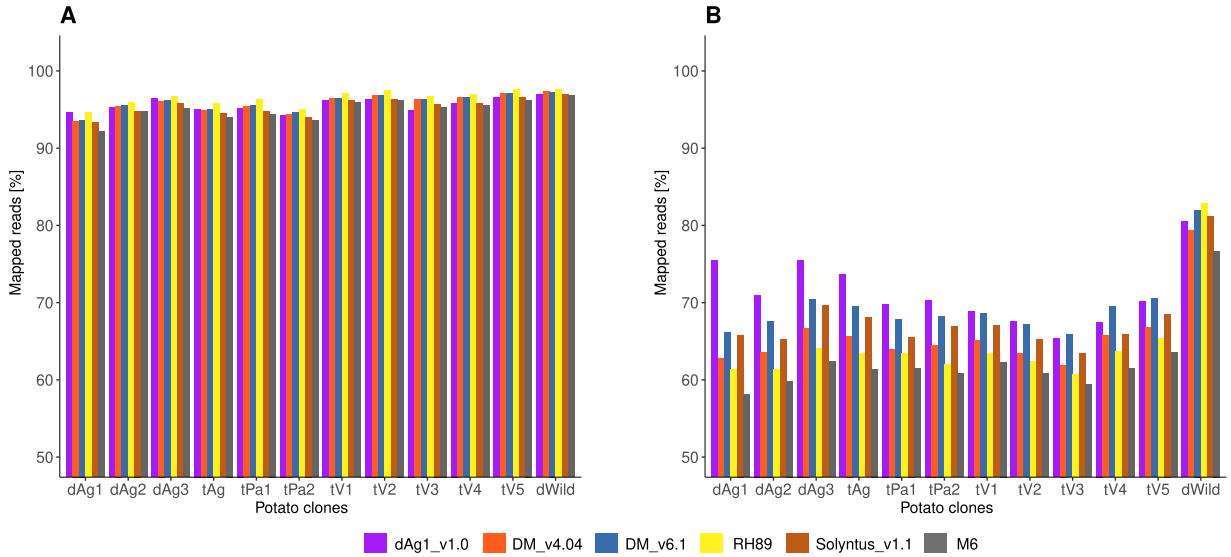

Figure 2 Percentage of 10xG linked reads of different potato clones mapped to different potato assemblies (A) and percentage of 10xG linked reads properly paired in mapping against different potato assemblies (B).

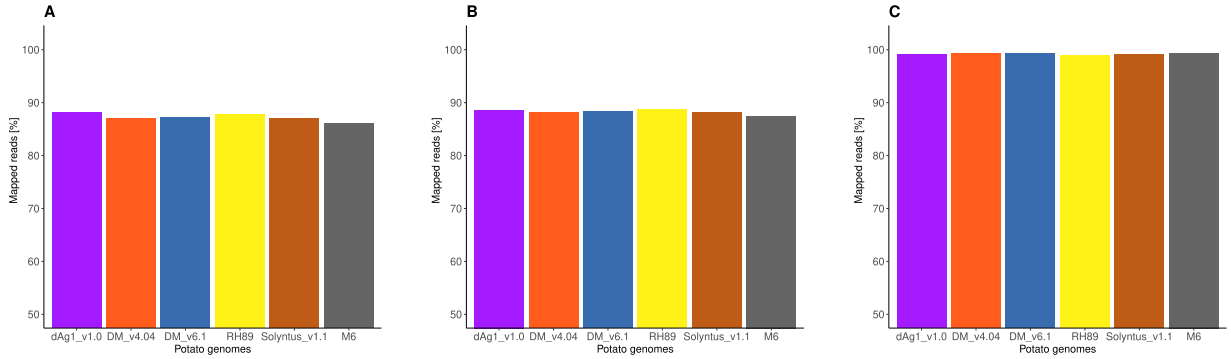

Figure 3 Percentage of dAg1 PacBio reads (A), tAg PacBio reads (B), and tAg high-quality Iso-seq RNA reads (C) mapped to different potato assemblies.

(Supplementary Figure S6). This illustrates the importance of alternative splicing and polyadenylation (Supplementary Figure S5B). The detailed analyses of these aspects using SQANTI3 did not reveal any systematic differences (Supplementary Figure S7) compared to that described for other plants (Abdel-Ghany *et al.* 2016; Wang *et al.* 2020) and are therefore not discussed further in detail.

Intragenomic diversity

We detected across the dAg1 genome 7,829,534 heterozygous SNV and indels which resulted in a sequence diversity between haplotypes of ~1% (Table 2). A similar amount of heterozygous variants was identified for the two other diploid Agria clones dAg2 and dAg3. These values are in the range of what was reported previously for diploid wild species *S. commersonii* (1.49%; Aversano *et al.* 2015) and M6 (0.68%; Leisner *et al.* 2018). In addition, 32,028 SV were detected between the two dAg1 haplotypes which are in the similar range of what was reported for RH89 (Zhou *et al.* 2020). For the three tetraploid clones tAg, tPa1, and tPa2, the frequency of heterozygous variants was with 3.1%, 3.6%, and 3.3%, respectively, about thrice higher than for the diploid ones. This is due to the fact that in tetraploid clones, more variants between haplotypes can occur compared to diploid clones, as more haplotypes are present. These results are in accordance with those of Hardigan *et al.* (2017), who found a similar relation between the variant frequencies of diploid and tetraploid clones which were 1.05% for diploid landraces and 2.73% for tetraploid cultivars.

The number of genes with at least one deleterious variant was assessed in our study. This number was for the diploid clones with values between 13,287 and 16,766 considerably higher than the deleterious mutations in 10,642 annotated genes described by Zhou *et al.* (2020) for the RH89 genome. This finding might be due to the usage of different approaches to detect deleterious mutations. In our study, short reads were mapped against the dAg1_v1.0 sequence, whereas Zhou *et al.* (2020) aligned the assembled chromosomes of RH89 to the DM_v4.03 sequence.

The number of genes with at least one deleterious variant observed for the tetraploid clones (tPa1 and tPa2) was with values of 27,927 and 28,357 about twice as high as for the diploid clones. Hence, the proportion of genes with at least one deleterious variant in 1-Mb windows across the genome is higher for tetraploids than for diploids (Figure 4). This might be due to that in tetraploid clones deleterious alleles can be more easily masked by non-deleterious alleles due to the higher number of alleles per gene. This explanation is supported by the higher number of genes with at

Table 2 Number of variants (SNV and indels) and genes with at least one deleterious variant among the haplotypes of a potato clone

Clone	Number of variants			Number of genes (del. variant)	
	Total	Heterozygous	Homozygous	Total	Homozygous
dAg1	7,829,534	7,829,534	—	13,287	—
dAg2	9,790,584	7,710,744	2,079,840	16,365	1,838
dAg3	9,461,662	7,975,910	1,485,752	16,766	1,436
tAg	25,559,532	25,495,186	64,346	26,134	25
tPa1	30,680,341	29,831,031	849,310	28,357	669
tPa2	28,666,770	27,156,995	1,509,775	27,927	1,060

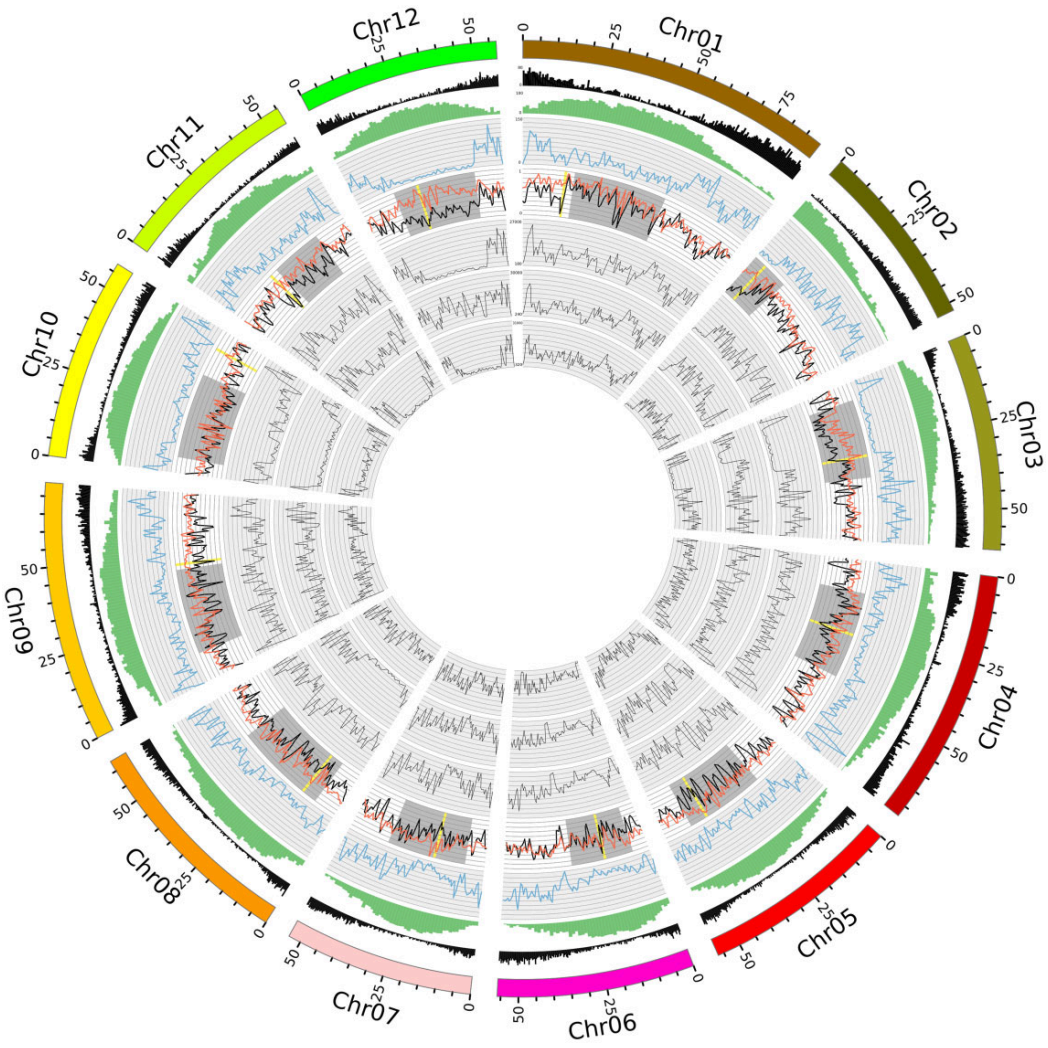

Figure 4 Distribution of genomic features across the potato genome. The outermost circle denotes the chromosome number and the physical position. The next inner circles report the distributions of genes (black), repeats (green) measured as percentage of masked bp, and structural variations (blue). The four most inner circles illustrate the proportion of genes with at least one deleterious variant in ∅dAg1-3 (black) and ∅tPa1-2 (orange), and heterozygous variants in dAg1, dAg2, and dAg3 in 1-Mb windows, respectively. The gray bars mark the pericentromeric regions, whereas the yellow bars mark the regions where the highest difference between the proportion of genes with at least one deleterious variant of ∅dAg1-3 and ∅tPa1-2 was identified.

least one homozygous deleterious variant for diploids (dAg2: 1,838; dAg3: 1,436) than for tetraploids (tPa1: 669; tPa2: 1,060). A similar number of genes with homozygous deleterious variants (1,753) was observed by Zhou *et al.* (2020). These findings indicate the higher masking potential of deleterious alleles in tetraploids, compared to diploids. Furthermore, our observation illustrates the high efforts that will be required to breed potato as a diploid hybrid crop. This is especially true as the proportion of the number of genes with at least one deleterious variant between ∅tPa1-2 and ∅dAg1-3 was significantly ($P < 0.001$, *t*-test, sample size:

Table 3 Q_{95}, median, and Q_5 of the block length in bp of phased variants

Clone	Q_{95} (bp)	Median (bp)	Q_5 (bp)
dAg1	626,568	6,824	6
dAg2	341,204	267	2
dAg3	251,607	301	2
tAg	1,207	188	16
tPa1	872	131	11
tPa2	822	116	9

Table 4 Percentage of phased blocks for which the haplotypes of progenies occurred in 0 to multiple copies in the parental clones

Samples	dAg1	dAg2	dAg3	tAg
		0/1/2 (%)		0/1/2/3/4 (%)
tAg	2.3/9.5/88.2	2.0/10.8/87.2	2.1/9.2/88.7	–
tPa1	2.1/11.8/86.1	2.9/11.2/85.9	3.3/11.2/85.5	1.8/5.0/14.5/23.7/55.0
tPa2	2.1/11.4/86.5	2.2/10.9/86.9	2.9/11.2/85.9	1.6/4.9/14.7/24.9/53.9

863) higher in pericentromeric regions, compared to subtelomeric regions. In the former, a purging of alleles is considerably more difficult due to the reduced recombination.

In addition to the frequency of sequence variants, the phasing of alleles is relevant to evaluate the possibilities of combining or separating alleles at neighboring loci by recombination. Recently, methods have been proposed for phasing that rely on long-read sequencing (*e.g.*, Schrinner *et al.* 2020). We have evaluated the use of linked-read sequencing for phasing the heterozygous variants for diploid and tetraploid clones. The resulting blocks of phased regions across the genomes had a median length of 116 bp for tPa2 and 6,824 bp for dAg1 (Table 3). The figures are discouraging with respect to the use of phasing information *e.g.*, in the context of genomic selection approaches (Stich and Van Inghelandt 2018). Nevertheless, these lengths are in accordance to the results of Yang *et al.* (2017) who phased the hexaploid sweet potato genome and obtained 542,361 phased regions, which covered about 30% of the genome.

Despite the short block length, the phased regions of parental and offspring clones were compared to each other with respect to the present alleles. In only 2.0–2.3% of the cases, the haplotypes (*i.e.*, phased variants) observed in the three diploid clones were not observed in tAg (Table 4). For the grandparents tPa1&2, these figures were with 2.1–3.3% slightly higher but still indicating a good phasing accuracy. More than 50% of all haplotypes of tAg were also observed in the four haplotypes of the parental clones tPa1&2. It was expected that two haplotypes of tAg would occur in tPa1 and the other two in tPa2. However, an in-depth evaluation of the phasing accuracy of tetraploids using 10xG linked reads was not possible due to the short phased regions.

Conclusions

In this study, we have created a chromosome-scale consensus sequence across the two haplotypes of a diploid clone derived from a tetraploid elite potato variety. This *de novo* assembly was performed with an optimal combination of today's sequencing technologies, comprising 10xG linked reads, PacBio long reads, and Hi-C reads. Comparisons of the new dAg1_v1.0 sequence to other potato genome assemblies pointed out the high divergence between the different potato clones and illustrated the potential of using dAg1_v1.0 sequence in breeding applications. The high amount of heterozygous SNV and indels, SV, and genes with at least one deleterious variant highlights the intragenomic diversity of the dAg1_v1.0 genome. Finally, in this study, we have shown that sequence variants of diploid potato clones could be phased using cost-efficient 10xG linked reads and the dAg1_v1.0 sequence. However, further improvements are needed to enlarge the phased regions to enable this approach in a breeding-related context.

Data availability

Supplementary File_S1 contains Supplementary Tables S1–S3 and Figures S1–S7. Supplementary Table S1 contains statistics of the sequencing technology data used in this study, including potato clones, data type, number of reads, median, Q_5, Q_{95}, and raw coverage. Supplementary Table S2 contains assembly statistics, namely number of contigs, assembly size, largest contig, N50, N90, L50, L90, Ns per 100 kb, and percentage of BUSCO genes, for an alternative assembly strategy. Supplementary Table S3 contains statistics of the different published potato assemblies with number of scaffolds, assembly size, assembly size considering only 12 chromosomes, scaffold N50, and Ns per 100 kb for 12 chromosomes. Supplementary Figure S1 shows the pipeline of the two different assembly strategies evaluated in this study. Supplementary Figure S2 shows the read depth and the percentage of masked bp in windows across all potato chromosomes. Supplementary Figure S3 depicts the alignment dot plot of dAg1_v1.0 and DM_v4.04 genomes. Supplementary Figure S4 represents the alignment dot plot of RH89 and DM_v6.1 genomes. Supplementary Figure S5 depicts the number of transcripts assigned to annotated and novel genes and alternative splicing events. Supplementary Figure S6 shows the number of isoforms per gene. Supplementary Figure S7 indicates the frequency distribution of Full Splice Match (FSM) transcripts. The Supplementary material is available via figshare repository (doi.org/10.6084/m9.figshare.14729943). Raw sequencing data of dAg1, dAg2, dAg3, and tAg have been deposited into the NCBI Sequence Read Archive (SRA) under the accession PRJNA729250. The genome sequence of dAg1_v1.0 and the corresponding annotation files are available via figshare repository (doi.org/10.6084/m9.figshare.14604780).

Acknowledgments

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

B.S. designed and coordinated the project; E.O.-H., J.R., S.H., K.M., B.T., A.R., V.P., J.B., and J.L. provided the genetic material; B.H. extracted DNA and prepared the libraries; R.F., M.W., R.G., and N.B. performed the analyses; R.F., M.W., R.G., and B.S. wrote the manuscript.

Funding

This study was funded by Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, Nordring-Kartoffelzucht- und Vermehrungs-GmbH, SaKa Pflanzenzucht GmbH & Co. KG, and the Federal Ministry of Food and Agriculture/Fachagentur Nachwachsende Rohstoffe (grantID 22011818, PotatoTools).

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 have read and approved the final manuscript.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, *et al.* 2016. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. 7:11706.

Altpeter F, Springer NM, Bartley LE, Blechl AE, Brutnell TP, *et al.* 2016. Advancing crop transformation in the era of genome editing. Plant Cell. 28:1510–1520.

Aversano R, Contaldi F, Ercolano MR, Grosso V, Iorizzo M, *et al.* 2015. The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. Plant Cell. 27:954–968.

Beddington J. 2010. Food security: contributions from science to a new and greener revolution. Philos Trans R Soc Lond B Biol Sci. 365:61–71.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 48:4.11.1–4.11.39.

Caruana BM, Pembleton LW, Constable F, Rodoni B, Slater AT, *et al.* 2019. Validation of genotyping by sequencing using transcriptomics for diversity and application of genomic selection in tetraploid potato. Front Plant Sci. 10:670.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44:e147.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, *et al.* 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 13:1050–1054.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, *et al.* 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 356:92–95.

Dudchenko O, Pham M, Lui C, Batra SS, Hoeger M, *et al.* 2018. Hi-C yields chromosome-length scaffolds for a legume genome, *Trifolium subterraneum*. Preprint at: 10.1101/473553

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, *et al.* 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3:95–98.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

FAO. 2008. Statistical data. Food and Agriculture Organization of the United Nations, Rome.

FAO. 2019. Statistical data. Food and Agriculture Organization of the United Nations, Rome.

Field MA, Rosen BD, Dudchenko O, Chan EK, Minoche AE, *et al.* 2020. Canfam-GSD: de novo chromosome-length genome assembly of the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. GigaScience. 9:1–12.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907

Ghurye J, Pop M, Koren S, Bickhart D, Chin CS. 2017. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 18:527.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8:1494–1512.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. BMC Bioinformatics. 19:348.

Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, *et al.*. 2017. Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc Natl Acad Sci U S A. 114:E9999–E10008. 10.1073/pnas.1714380114 29087343

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, *et al.* 2018. Tigmint: correcting assembly errors using linked reads from large molecules. BMC Bioinformatics. 19:393.

Jansky S, Navarre R, Bamberg J. 2019. Introduction to the special issue on the nutritional value of potato. Am J Potato Res. 96:95–97.

Jiang T, Liu YY, Jiang Y, Li J, Gao Y, *et al.* 2020. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 21:189.

Jiao WB, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 36:64–70.

Kadota M, Nishimura O, Miura H, Tanaka K, Hiratani I, *et al.* 2020. Multifaceted Hi-C benchmarking: what makes a difference in chromosome-scale genome scaffolding? Gigascience. 9:1–15.

Kinkar L, Gasser RB, Webster BL, Rollinson D, Littlewood DTJ, *et al.* 2021. Nanopore sequencing resolves elusive long tandem-repeat regions in mitochondrial genomes. Int J Mol Sci. 22:1811.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* 2017. Secure because math: a deep-dive on machine learning-based monitoring. Genome Res. 27:722–736.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5: 59.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, *et al.* 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47: D807–D811.

Kuderna LF, Lizano E, Julià E, Gomez-Garrido J, Serres-Armero A, *et al.* 2019. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nat Commun. 10:4.

Kyriakidou M, Anglin NL, Ellis D, Tai HH, Strömvik MV. 2020. Genome assembly of six polyploid potato genomes. Sci Data. 7:88.

Leisner CP, Hamilton JP, Crisovan E, Manrique-Carpintero NC, Marand AP, *et al.* 2018. Genome sequence of M6, a diploid inbred clone of the high glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense* reveals residual heterozygosity. Plant J. 94:562–570.

Lenaerts B, Collard BC, Demont M. 2019. Review: improving global food security through accelerated plant breeding. Plant Sci. 287: 110207.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.*; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Lieberman-Aiden E, Berkum NLV, Williams L, Imakaev M, Ragoczy T, *et al.* 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 326:289–294.

Liu C, Cheng YJ, Wang JW, Weigel D. 2017. Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. Nat Plants. 3:742–748.

Liu C, Weigel D. 2015. Chromatin in 3D: progress and prospects for plants. Genome Biol. 16:170.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, *et al.* 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat Commun. 10:260.

Manrique-Carpintero NC, Coombs JJ, Veilleux RE, Buell CR, Douches DS. 2016. Comparative analysis of regions with distorted segregation in three diploid populations of potato. G3 (Bethesda). 6: 2617–2628.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, *et al.* 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 14:e1005944.

Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, *et al.* 2018. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. Nature. 563:501–507.

Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, *et al.* 2016. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. Biotechniques. 61:203–205.

Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, *et al.* 2020. Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience. 9:1–11.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44:e113.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19:460.

Schrinner S, Mari RS, Ebler J, Rautiainen M, Seillier L, *et al.* 2020. Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. 21:252.

Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, *et al.* 2013. Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 (Bethesda). 3:2031–2047.

Shearman JR, Sonthirod C, Naktang C, Sangsrakru D, Yoocha T, *et al.* 2020. Assembly of the durian chloroplast genome using long PacBio reads. Sci Rep. 10:15980.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013–2015. <http://www.repeatmasker.org>

Spooner DM, McLean K, Ramsay G, Waugh R, Bryan GJ. 2005. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc Natl Acad Sci U S A. 102: 14694–14699.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 24:637–644.

Stich B, Van Inghelandt D. 2018. Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. Front Plant Sci. 9:159.

Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, *et al.* 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28: 369–411.

Thankaswamy-Kosalai S, Sen P, Nookaew I. 2017. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics. 109:186–191.

The UniProt Consortium 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47:D506–D515.

Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, *et al.* 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 19:40.

Uitdewilligen JG, Wolters AM, D'hoop BB, Borm TJ, Visser RG, *et al.* 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One. 8:e62355.

van Lieshout N, van der Burgt A, de Vries ME, ter Maat M, Eickholt D, Esselink D, *et al.* 2020. Solyntus, the new highly contiguous reference genome for potato (*Solanum tuberosum*). G3 (Bethesda). 10: 3489–3495.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. Nat Protoc. 11:1–9.

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, *et al.* 2020. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann Hum Genet. 84:125–140.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.

Wang B, Tseng E, Baybayan P, Eng K, Regulski M, *et al.* 2020. Variant phasing and haplotypic expression from long-read sequencing in maize. Commun Biol. 3:78.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, *et al.* 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience. 4:35.

Xu X, Pan S, Cheng S, Zhang B, Mu D, *et al.*; Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. Nature. 475:189–195.

Yang J, Moeinzadeh MH, Hu F, Boerno S, Sun Z, *et al.* 2017. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nat Plants. 3:696–703.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 34:725–731.

Zhang C, Wang P, Tang D, Yang Z, Lu F, *et al.* 2019. The genetic basis of inbreeding depression in potato. Nat Genet. 51: 374–378.

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, *et al.* 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 34:303–311.

Zhou Q, Tang D, Huang W, Yang Z, Zhang Y, *et al.* 2020. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Genet. 52:1018–1023.

Communicating editor: G. Morris

SUPPLEMENTARY INFORMATION

10xG assembly as backbone

Two 10xG libraries (1 and 2) of dAg1 were used for *de novo* assemblies using supernova (v2.1.1) (Weisenfeld et al., 2017) to generate pseudo-haploid assemblies. Based on the raw assembly statistics, we decided to use the assembly of the 10xG library 2 for the next steps. PacBio reads were used to scaffold the 10xG assembly using SSPACE-LongRead (v1.1) (Boetzer and Pirovano, 2014) in four iterations.

This assembly strategy has as backbone a 10xG assembly with 59,831 scaffolds and an N50 of 0.306 Mb (Supplementary Table S2). This assembly was scaffolded in four rounds with SSPACE-LongRead (v1.1) (Boetzer and Pirovano, 2014) and PacBio reads. After the fourth round, N50 had triplicated to 0.901 Mb and the number of contigs decreased to 15,034. The proportion of complete gene orthologs from the Solanaceae set, in the following designated as BUSCO statistics, improved slightly from 93% of the 10xG assembly to 94% after scaffolding.

REFERENCES

Boetzer M, Pirovano W (2014), SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15:211

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

Table S1: Potato clones and sequencing data used in this study.

Clone	Data type	# reads [M]	Median [bp]	Q_{95} [bp]	Q_5 [bp]	Raw coverage
dAg1	PacBio	13.9	9,671	16,896	431	142x
dAg1	10xG library1	157.1	2x150	-	-	56x
dAg1	10xG library2	318.6	2x150	-	-	114x
dAg1	Hi-C	129.4	2x150	-	-	45x
dAg2	10xG library1	182.6	2x150	-	-	60x
dAg3	10xG library1	168.7	2x150	-	-	57x
tAg	Iso-seq	0.212	2,895	4,753	1,385	-
tAg	PacBio	7.3	7,242	28,213	339	85x
tAg	10xG library1	195.3	2x150	-	-	66x
tAg	10xG library2	58.02	2x150	-	-	20x
tAg	10xG library2b	265.9	2x150	-	-	90x
tAg	10xG library3	190.5	2x150	-	-	65x
tPa1	10xG library1	161.7	2x150	-	-	55x
tPa1	10xG library2	183.9	2x150	-	-	62x
tPa2	10xG library1	187.9	2x150	-	-	64x
tPa2	10xG library2	184.7	2x150	-	-	63x
tV1	10xG library1	538.4	2x150	-	-	184x
tV2	10xG library1	555.1	2x150	-	-	190x
tV3	10xG library1	551.3	2x150	-	-	188x
tV4	10xG library1	352.5	2x150	-	-	120x
tV5	10xG library1	743.0	2x150	-	-	254x
dW[1]	Illumina	157.1	2x150	-	-	57x

[1]Kyriakidou et al. (2020)

Table S2: Assembly statistics of alternative genome assembly strategy for dAg1. For details see Supplementary Information.

Assembly step	#Contigs	Assembly size [Mb]	Largest contig [Mb]	N50 [Mb]	N90 [Mb]	L50	L90	Ns per 100kb	BUSCO [%]
Assembling									
supernova	59,831	1048.7	7.002	0.306	0.006	704	19,623	9,191	93
SSPACE-LongRead 1x	45,626	1066.2	7.014	0.350	0.008	643	12,043	10,517	93
SSPACE-LongRead 2x	26,340	1099.9	7.730	0.503	0.024	499	4,945	12,951	94
SSPACE-LongRead 3x	17,608	1121.2	9.209	0.689	0.057	384	2,552	14,519	94
SSPACE-LongRead 4x	15,034	1127.4	9.451	0.901	0.090	310	1,757	14,993	94

Table S3: Comparison of the assembly statistics of different potato genomes.

Genome	#Scaffolds	Assembly size [Mp]	Assembly size (12 Chromosomes) [Mb]	Scaffold N50 [Mb]	Ns per 100kb (12 Chromosomes)
dAg1_v1.0	12+614	812.070	750.058	57.412	788
DM_v4.04[1]	12+14,841	884.168	725.017	1.345	12,527
DM_v6.1[1]	12+276	741.585	731.288	59.671	11
Solyntus_v1.1[2]	12	716.171	716.171	63.702	1
RH89[*,3]	24+3,125	1695.610	1664.030	1.743	1,278
M6[4]	12+8,258	825.768	499.048	0.714	2,510

[*]Diploid data,[1]Pham et al. (2020), [2]van Lieshout et al. (2020), [3]Zhou et al. (2020),

[4]Leisner et al. (2018)

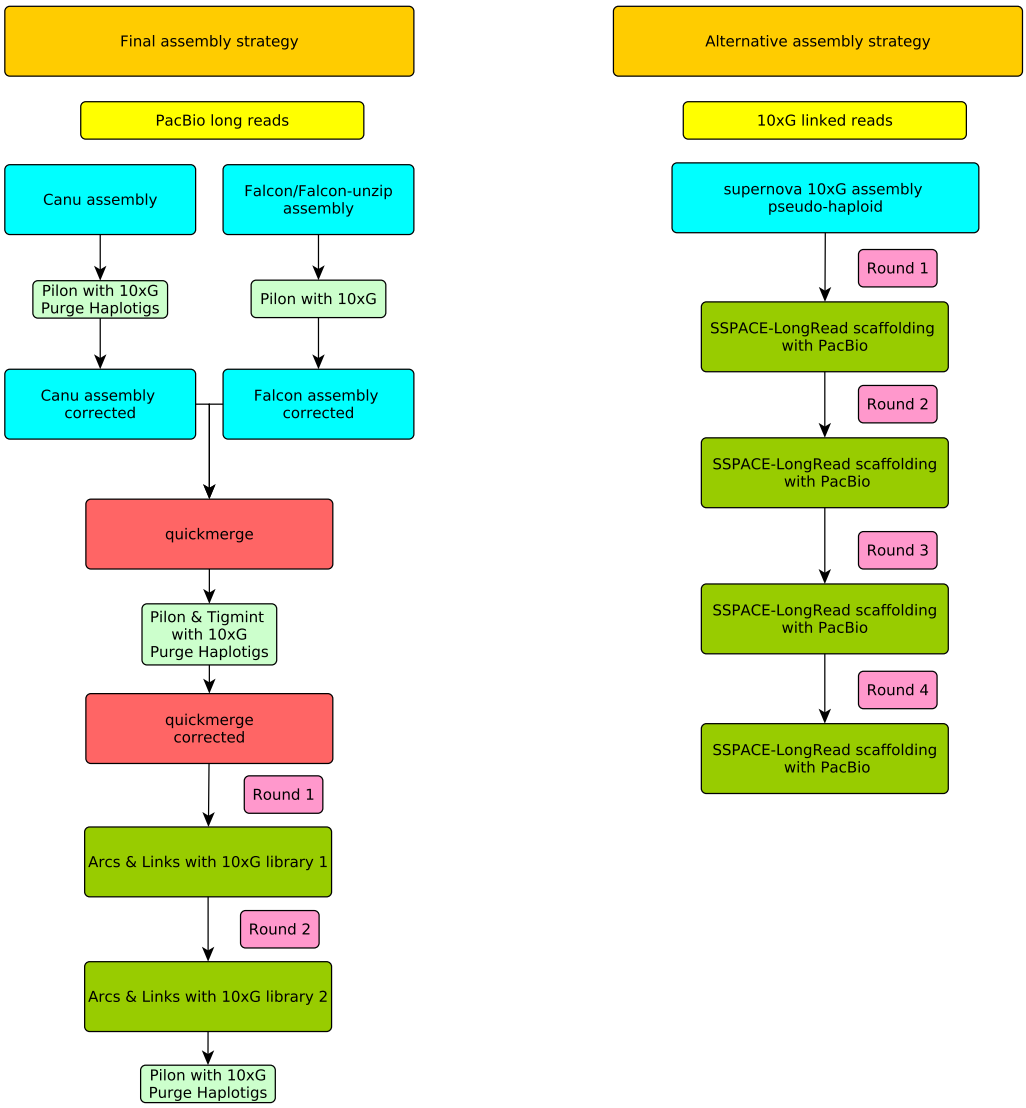

Fig. S1: Graphical illustration of the evaluated genome assembly strategy.

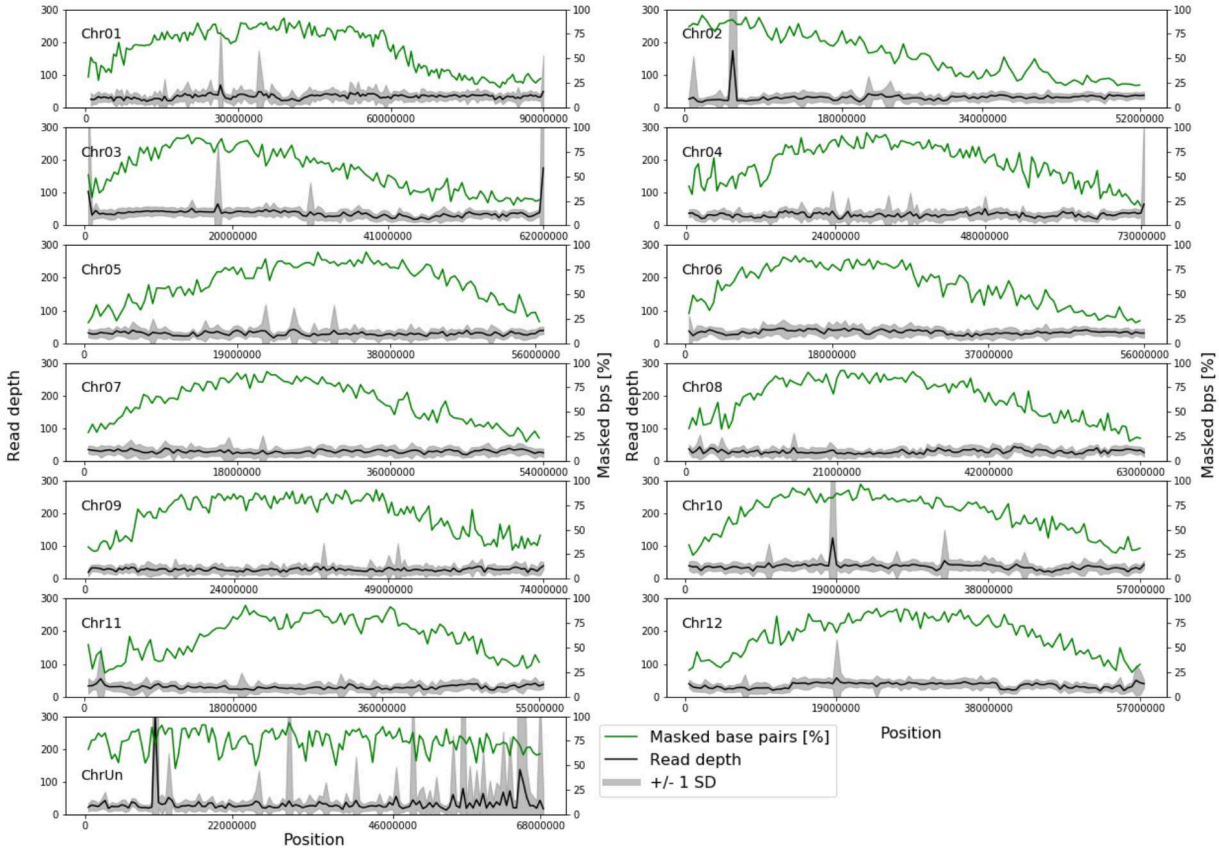

Fig. S2: Read depth (left axis) and percentage of masked bp (right axis) averaged across 500kb windows across all chromosomes.

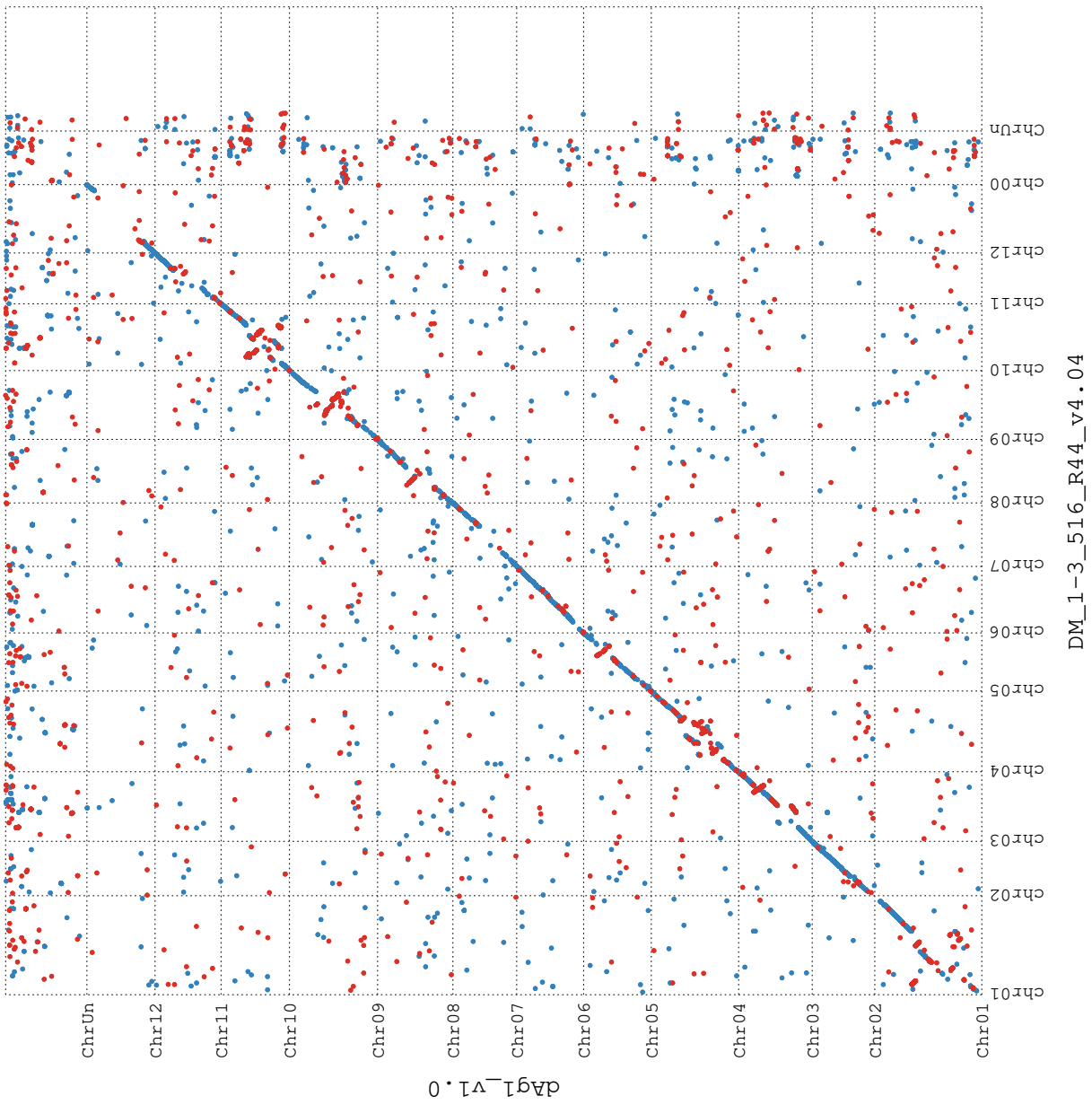

Fig. S3: Alignment dot plot of dAg1_v1.0 (vertical) and DM_v4.04 genomes (horizontal). Each dot indicates an alignment with a length $\geq$ 1000bp between the two genomes. Forward and reverse alignments are represented as blue and red dots, respectively.

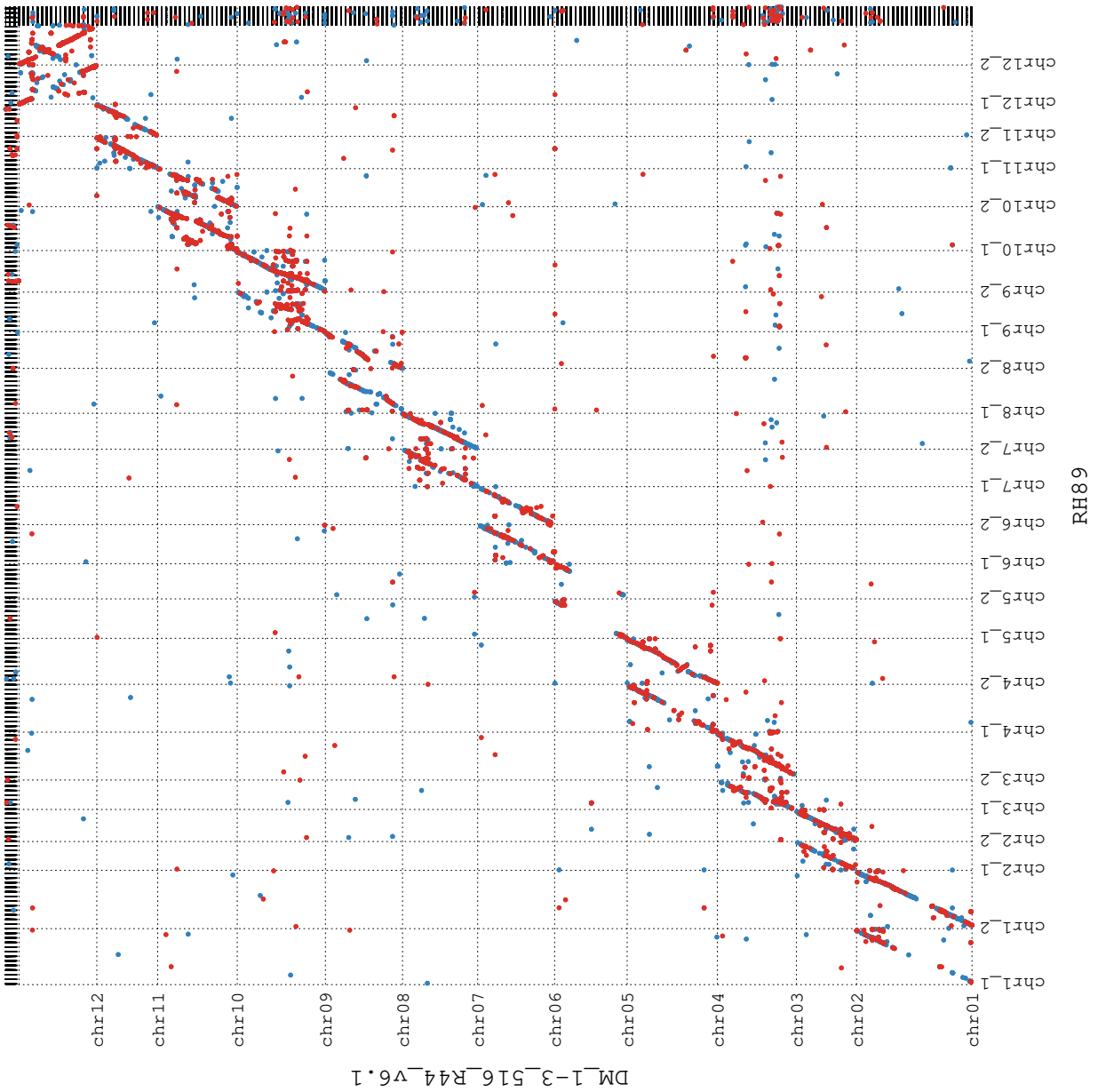

Fig. S4: Alignment dot plot of RH89 (vertical) and DM_v6.1 genomes (horizontal). Each dot indicates an alignment with a length $\geq$ 1000bp between the two genomes. Forward and reverse alignments are represented as blue and red dots, respectively.

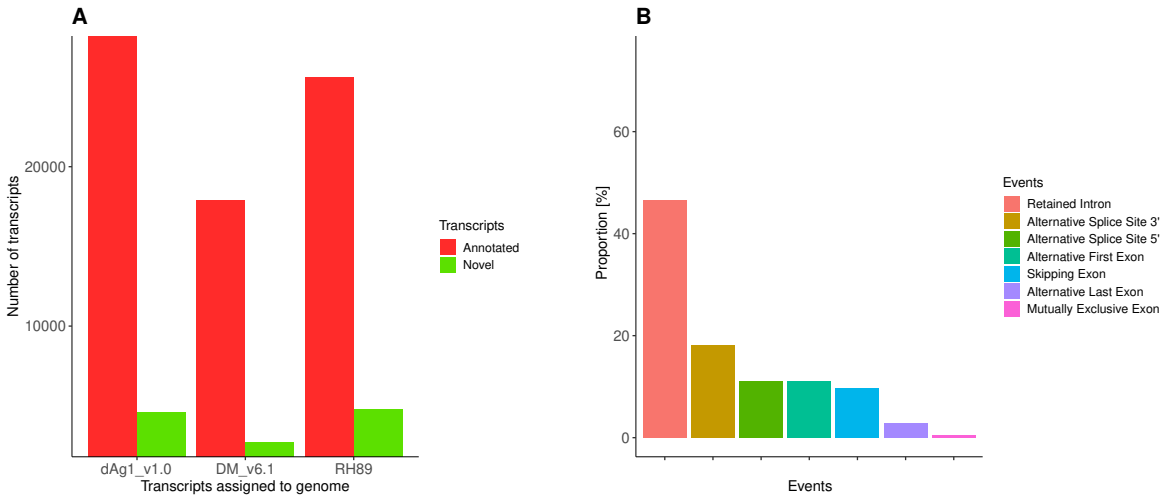

Fig. S5: Number of transcripts assigned to annotated and novel genes after mapping to three different potato assemblies (A) and alternative splicing events detected in the analysis of transcripts (B) of tAg against dAg1_v1.0 genome.

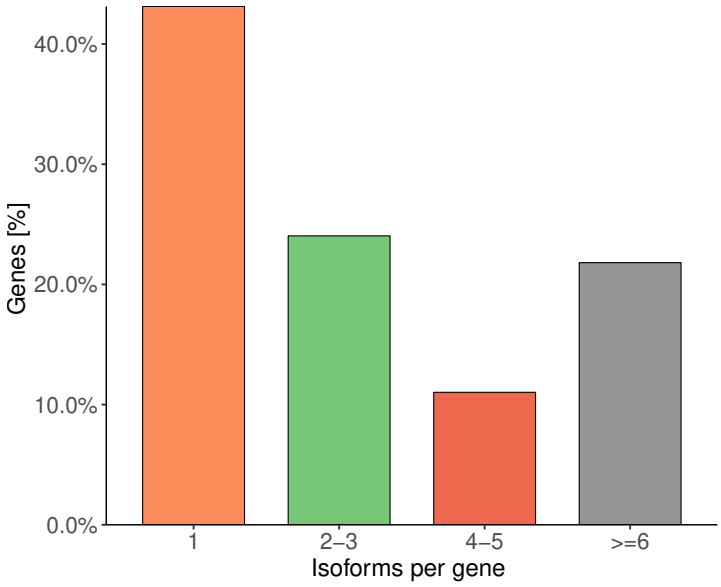

Fig. S6: Histogram of the number of isoforms per gene.

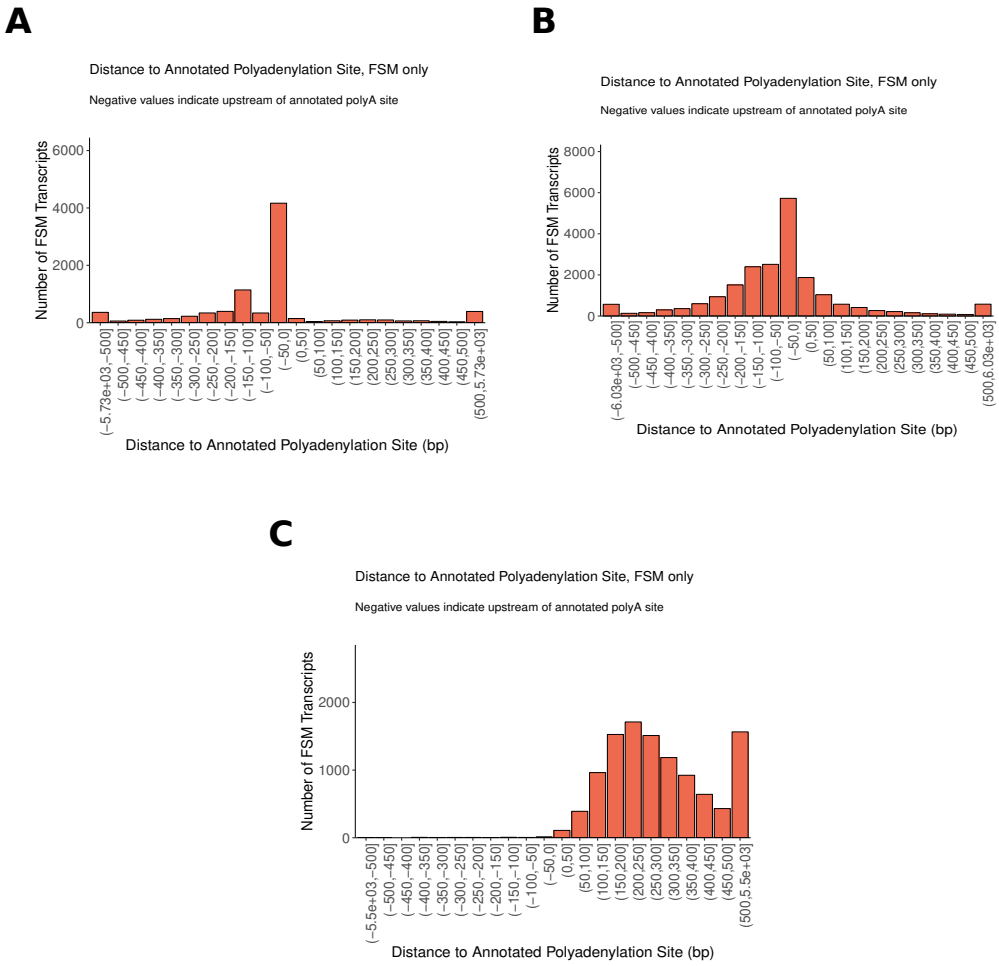

Fig. S7: Frequency distribution of Full Splice Match (FSM) transcripts as a function of the binned distance between Iso-seq transcript poly(A) site and the corresponding annotated poly(A) site of dAg1_v1.0 (A), DM_v6.1 (B), and RH89 (C) annotation sets.

5. Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits

This manuscript was published in BMC Genomics in October, 2019.

Authors:

Marius Weisweiler, Amaury de Montaigu, David Ries, Mara Pfeifer, Benjamin Stich

Contribution: Shared first author

Benjamin Stich designed and coordinated the project.

Marius Weisweiler, Amaury de Montaigu, David Ries, and Benjamin Stich performed the data analyses.

Marius Weisweiler, Amaury de Montaigu, and Benjamin Stich wrote the manuscript.

Amaruy de Montaigu collected the materials and prepared and purified mRNA and DNA samples.

Weisweiler *et al. BMC Genomics* (2019) 20:787
https://doi.org/10.1186/s12864-019-6174-3

BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits

Marius Weisweiler[1†], Amaury de Montaigu[1†], David Ries[1], Mara Pfeifer[1] and Benjamin Stich[1,2*]

Abstract

Background: Barley is the world's fourth most cultivated cereal and is an important crop model for genetic studies. One layer of genomic information that remains poorly explored in barley is presence/absence variation (PAV), which has been suggested to contribute to phenotypic variation of agronomic importance in various crops.

Results: An mRNA sequencing approach was used to study genomic PAV and transcriptomic variation in 23 spring barley inbreds. 1502 new genes identified here were physically absent from the Morex reference sequence, and 11,523 previously unannotated genes were not expressed in Morex. The procedure applied to detect expression PAV revealed that more than 50% of all genes of our data set are not expressed in all inbreds. Interestingly, expression PAV were not in strong linkage disequilibrium with neighboring sequence variants (SV), and therefore provided an additional layer of genetic information. Optimal combinations of expression PAV, SV, and gene abundance data could enhance the prediction accuracy of predicting three different agronomic traits.

Conclusions: Our results highlight the advantage of mRNA sequencing for genomic prediction over other technologies, as it allows extracting multiple layers of genomic data from a single sequencing experiment. Finally, we propose low coverage mRNA sequencing based characterization of breeding material harvested as seedlings in petri dishes as a powerful and cost efficient approach to replace current single nucleotide polymorphism (SNP) based characterizations.

Keywords: Barley, Multi-tissue transcriptomics, Presence/absence variation, *Omic prediction, Genomic selection

Background

A priority of modern agriculture is to increase the productivity of crops to meet the demands of a growing human population. The urge of achieving significant yield gains is amplified by the current context of climate change, competition for land, and limited natural resources [1]. Plant genetics and breeding are considered among the disciplines that have the highest potential to tackle this challenge. One of the major approaches used in plant breeding to increment yield gains is to exploit the natural genetic variation present in the crop species' gene pool. Barley was domesticated more than 10,000 years ago in the fertile crescent [2]. Its cultivation area has progressively expanded to a wide range of latitudes, and it is now the fourth most important cereal in the world [3]. Barley has also become an important model cereal species for research, partly because its tolerance to stress surpasses that of other major crops including wheat and rice [4]. Moreover, the diploid genome of barley facilitates genetics studies.

To exploit the natural genetic variation present in the gene pool of barley, genomic tools such as single

*Correspondence: benjamin.stich@hhu.de
†Marius Weisweiler and Amaury de Montaigu contributed equally to this work.
[1]Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany
[2]Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany

nucleotide polymorphism (SNP) arrays have been developed [5]. The availability of a reference genome sequence facilitates the use of next generation sequencing technologies for the discovery of novel sequence variants [6]. This allowed e.g. to characterize most of the barley accessions of the German ex situ genebank using a genotyping by sequencing approach [7].

Genome wide quantification of gene expression has also been accessible in barley since many years through the development of gene expression arrays [8]. This technology allowed addressing how the barley transcriptome varied between tissues [9], and how it responded to pathogens and to environmental cues such as vernalization and heat [10–13]. eQTL studies with these arrays further revealed a complex pattern of genome-wide regulation of barley genes [14], and described how limited pleiotropy acted on gene expression in a tissue dependent manner [15]. With the release of a high quality reference sequence, resequencing technologies are successfully providing novel information on the barley genome and transcriptome that had remained inaccessible [16–18].

It is now accepted that a significant proportion of the genes of plant genomes are not expressed (expression presence/absence variation; ePAV) or are even completely absent (genomic PAV; gPAV) in subsets of genotypes, and make up what is known as the dispensable transcriptome or genome [19–21]. The occurrence of PAV in crops has extensively been reported for maize and rice [19–25]. However, up to now, little information is available concerning the extent and distribution of PAV in the barley genome [18, 26].

Prediction of phenotypic variation in the context of genomic selection, which is nowadays an essential component of plant breeding programs, is performed based on SNP genotyping profiles. Previous studies on the use of metabolome and lipidome variation to predict phenotypic traits of maize revealed high but lower prediction accuracies compared to SNP information [27, 28]. Only the use of microarray based transcriptome information for prediction of phenotypic traits in maize resulted for a subset of the traits in increased prediction accuracies especially when combined with SNP genotyping information [29]. However, the transcriptomic characterization of genotypes by mRNA sequencing has the advantage that also SNP information can be extracted from such a data set. In addition, the cost of characterizing genetic material by mRNA sequencing can be influenced by modifying the sequencing depth. Despite these advantages, no earlier study examined the prediction accuracies of predictors extracted from mRNA sequencing data sets. Furthermore, an evaluation of the prediction accuracy of PAV has to our knowledge not yet been performed, despite that single PAV have been shown to contribute to phenotypic variation of selected traits in various crops (for review see Gabur et al. [30]).

In this study, we explored the genomic and transcriptomic landscape of 23 spring barley landraces and cultivars which were selected based on their genetic and phenotypic diversity as parents of a joint linkage and association mapping population. The objectives of our study were to (i) characterize genomic and transcriptomic variation in the barley genome using multi-tissue mRNA sequencing, (ii) assess the proportion of ePAV that are due to gPAV, (iii) examine how accurately the different layers of genomic and transcriptomic variation predict phenotypic variation of various agronomic traits.

Results

To study genomic diversity in spring barley inbreds, we first selected 23 inbreds from a panel of 224 representing a broad range of origins [31] (Additional file 1: Table S1). mRNA was extracted from seedlings and leaves of all of these 23 inbreds, and from apex of a subset of six inbreds (Additional file 1: Table S1). Gene expression was determined for each individual sample by sequencing the mRNA. Out of the 73,187 expressed genes across seedlings, leaves, and apex samples, 11,523 genes mapped to regions of the Morex reference genome where no gene had previously been annotated (Additional file 1: Figure S1). We considered a gene as newly annotated gene if it was detected in at least two samples. A total of 3,482 genes mapped to the unknown chromosome of the Morex reference sequence, where 581 of these were newly annotated genes. The average length of the newly annotated genes was 5,470 bp.

We additionally identified 1,502 new contigs, with an average gene length of 494 bp, that did not map to any of the seven barley chromosomes. These contigs were designated in the following as newly identified genes, although a portion of these contigs might not actually be protein coding genes, if they were expressed in at least two samples, and if they showed homology to at least one gene of one out of eight plant species. In total, 96% of the homologous genes were found in other cereals of the *Triticacea* tribe but not in more distantly related species (Fig. 1A), indicating that they were not conserved across the plant kingdom but might fulfill functions specific to barley and closely related species. In addition, only 280 of the newly identified genes had an unknown gene annotation compared to the eight plant species. Altogether, 67% of the newly identified genes were expressed in all three tissues (Fig. 1B), making it unlikely that they are due to technical artifacts. We next tested if the newly identified genes were found predominantly in isolated inbreds, or if their presence was common across our set of spring barley accessions. This analysis revealed that about 25% of newly

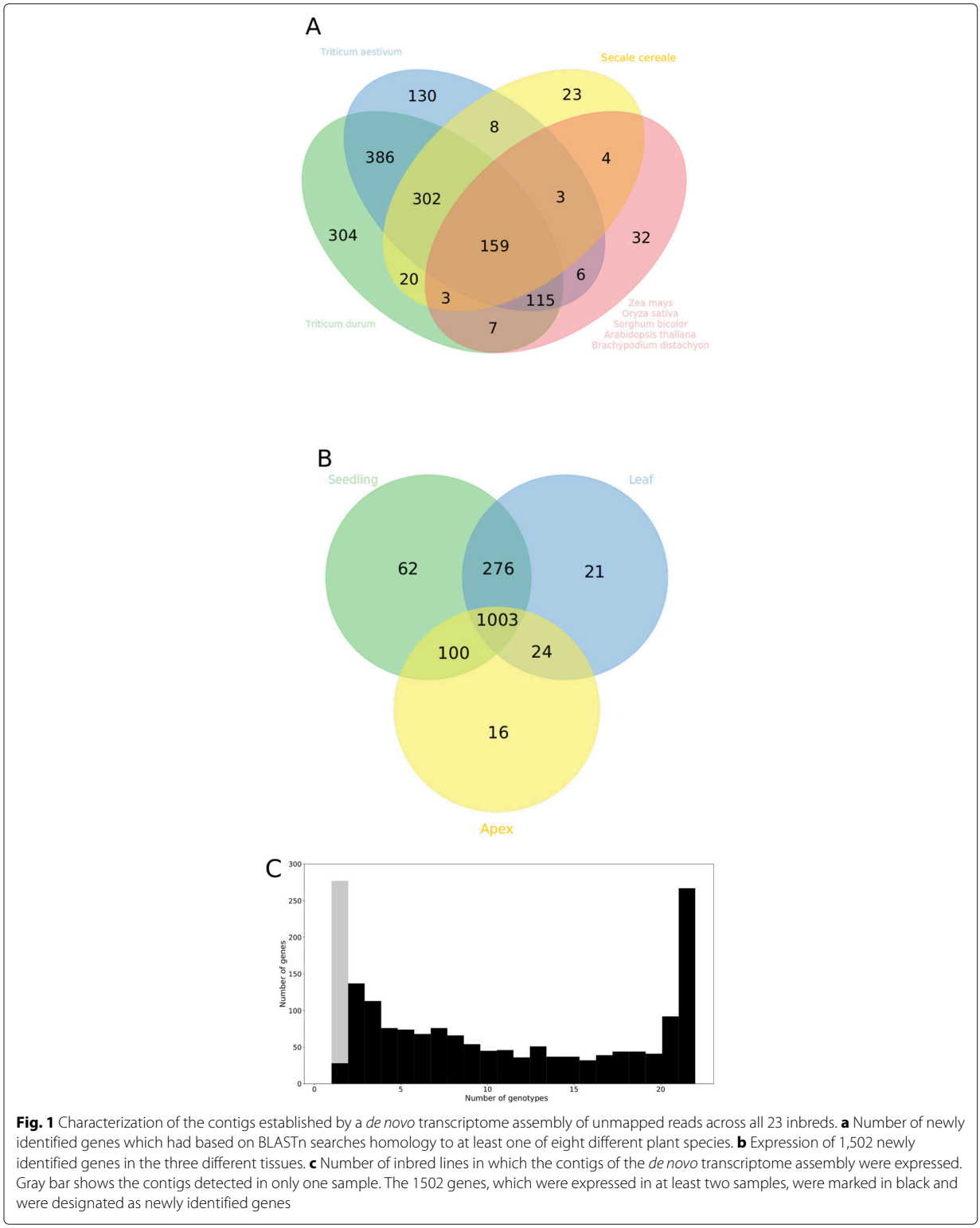

Fig. 1 Characterization of the contigs established by a *de novo* transcriptome assembly of unmapped reads across all 23 inbreds. **a** Number of newly identified genes which had based on BLASTn searches homology to at least one of eight different plant species. **b** Expression of 1,502 newly identified genes in the three different tissues. **c** Number of inbred lines in which the contigs of the *de novo* transcriptome assembly were expressed. Gray bar shows the contigs detected in only one sample. The 1502 genes, which were expressed in at least two samples, were marked in black and were designated as newly identified genes

identified genes were expressed in isolated inbreds, and another 25% in all inbreds (except Morex, Fig. 1C).

The high number of genes absent in Morex inspired us to systematically explore ePAV among the 23 inbreds. ePAV were defined as genes whose expression was detected/not detected in at least two inbreds. A total of 38,810 barley genes were detected as ePAV, of which 28,340 had previously been annotated in the reference genome (Additional file 1: Table S2). ePAV were enriched in genes implicated in very diverse biological processes (Additional file 1: Table S3). The average length of ePAV (4162 bp) was significantly shorter than that of non-ePAV genes (9458 bp). In contrast, the average coding sequence length of ePAV (411 bp) was longer than that of non-ePAV genes (282 bp). Non-ePAV genes only rarely corresponded to newly identified genes or newly annotated genes (Fig. 2), and in fact, 80.6% of the newly annotated genes and 78.8% of the newly identified genes were also detected as ePAV (Additional file 1: Table S2). ePAV were significantly ($P < 0.05$) unevenly distributed along the chromosomes, with the highest frequency of occurrence close to the centromers (Fig. 3).

The robustness of our ePAV detection procedure was evaluated using a resampling simulation. In 50 replications, 20% of the gene length of each gene was used for transcript calling and ePAV detection. Across the 50 replications, the average number of genes as well

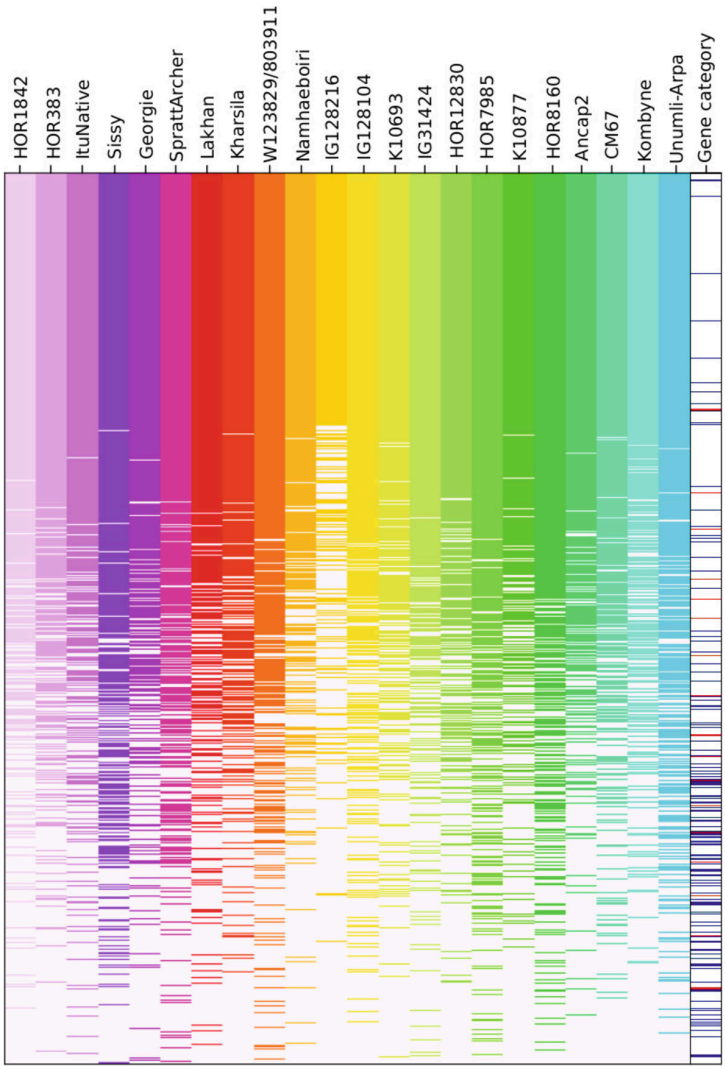

Fig. 2 Gene expression of all inbreds. Presence and absence of the 73,187 genes across all inbreds. Genes were sorted according to their presence across all inbreds (top to bottom). Presence of a gene is highlighted as colored and absence as white bar. The last column illustrates the three different gene categories as white (IBSC), blue (newly annotated), and red (newly identified) bars

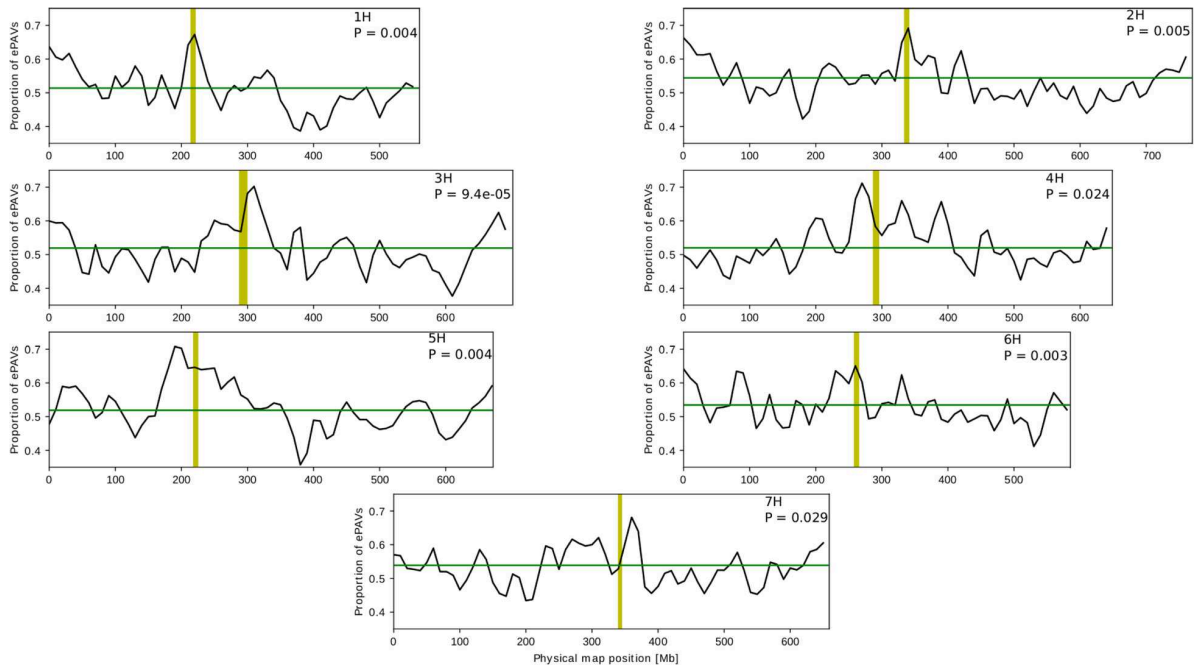

Fig. 3 Distribution of expression presence/absence variation (ePAV) across the physical map of the barley chromosomes. Proportion of ePAV among all genes within sliding windows of 20 Mb is given on the y-axis. The green line shows the average proportion of ePAV among all genes of a chromosome and the P value indicates, if the distribution deviates significantly from an uniform distribution. The yellow line illustrates the centromeric region

as the number of detected ePAV was about 67,000 and 35,400, respectively, only slightly lower than the 73,187 and 38,810 detected when considering the entire gene length (Additional file 1: Table S2).

Information on the proportion of ePAV that are due to gPAV is generally scarce. In order to estimate it, we used SNP genotyping profiles of segregating populations. The 23 inbreds had previously been crossed following a double round robin design [32] to generate a joint linkage and association mapping population. SNP genotyping profiles obtained with a 50K SNP array were available for these 45 populations. We searched for SNP for which missing data were segregating as a monogenic character, and used this pattern to assign presence/absence calls to the parental inbreds. Using such SNP located within genes, which we refer to as gPAV-SNP, we calculated the proportion of gPAV-SNP that were also detected by our procedure as ePAV, and considered this value as an estimation of the power to detect gPAV by our ePAV detection procedure (Additional file 1: Figure S2).

Based on the criterion that a gene is considered an ePAV if it has a present and an absent call in at least two inbreds, the power of gPAV detection was 34.6% (Table 1). This means that out of all gPAV, which we detected based on gPAV-SNP from the SNP array data, we identified 34.6% of it as ePAV in our mRNA sequencing data. It could be possible that parts of genes are still expressed even though a small fragment of their sequence was deleted. In this case, the genes would have a presence call though the region around the gPAV-SNP were

Table 1 Detection procedure of presence/absence variation

		Expression of gene			Expression of ±5bp genic SNP		
Tissue	*t*	1-β^*	α^*	*o*	1-β^*	α^*	*o*
Leaf&Seedling&Apex	1	45.0	88.8	88.6	64.2	90.8	85.3
	2	34.6	87.5	81.8	53.0	90.1	81.8
	3	30.1	86.7	79.5	44.8	90.0	79.5
	4	25.1	87.0	77.3	38.4	89.8	77.3
	5	21.4	86.5	77.3	32.7	89.7	77.3
Leaf&Seedling	2	35.2	87.5	81.8	52.4	90.3	79.5
Leaf	2	34.0	89.0	73.8	45.9	91.2	72.2
Seedling	2	28.1	86.7	76.7	45.9	90.5	76.2

Statistical power ($1 - \beta^*$) and the empirical type I error rate (α^*) to detect genomic presence/absence variation (gPAV) by expression PAV (ePAV), where *t* is the minimum number of inbreds that must have a present and absent call for a gene, *o* the percentage of common presence/absence values across all inbreds between ePAV and the genic PAV-SNP. We considered two scenarios: (i) the expression across the entire gene or (ii) the expression determined in a 10 bp window around the genic SNP was used to determine ePAV.

not covered by reads. For this reason, we also calculated the power of our procedure when detecting ePAV exclusively based on the 10bp sequence window surrounding the gPAV-SNP instead of using FPKM-values for the entire gene sequences. In this case, the power increased to 53%. The empirical type I error rate, defined as the proportion of ePAV that are not gPAV, was about 90%. Finally, the similarity between presence/absence patterns in the 23 inbreds of ePAV and gPAV-SNP was very high, ranging from 70% to 90%.

We were interested in knowing how independent the detected ePAV are from the local genomic pattern. First, the mRNA sequencing data was used to call sequence variants (SV) within exon sequences. A total of 133,566 SV were detected. We then determined the extent of linkage disequilibrium (LD) between each ePAV and neighboring SV located within 100 kb. Only 17.5% of all ePAV have at least one SV within 100 kb that has an $r^2 \geq 0.4$ (Table 2). This figure is even lower than for SV that are located outside the 100 kb window. In contrast, more than 85% of SV that are neighboring an ePAV show an $r^2 \geq 0.4$ with another SV within 100 kb. Therefore, ePAV provide an additional layer of genetic information compared to SV. This idea was confirmed by comparing principal component analyses (PCA) performed based on SV and ePAV. Both PCA revealed the existence of two clusters of inbreds defined by the row type of the inbreds (Additional file 1: Figure S3). Principal components 1 from both PCAs were significantly correlated with each other ($r^2 = 0.4928709$, $p = 0.0002706$), and a similar result was observed for principal components 2 ($r^2 = 0.3980411$, $p = 0.001643$). However, these analyses also reveal that the relationship of the inbreds within clusters differs between the two sources of molecular variation. A similar trend was observed when comparing the transcriptomic variation (T) with that of ePAV and SV. Mantel tests of distance matrices calculated from T, SV, and ePAV data indicated only significant correlations between the seedling transcriptome and SV ($r = 0.2581$, $p = 0.03969$).

Table 2 Linkage disequilibrium between expression presence/absence variation (ePAV) and sequence variants (SV)

r^2	[1.0,0.8]	(0.8,0.6]	(0.6,0.4]	(0.4,0.2]	(0.2,0]
	Percentage of r^2_{max} between ePAV and SV				
linked	0.0	0.0	17.5	49.9	31.1
unlinked	0.0	0.1	23.4	54.0	22.1
	Percentage of r^2_{max} between closest SV beside ePAV and SV				
linked	0.0	34.1	52.9	12.9	0.0
unlinked	0.0	21.1	52.7	25.2	0.0

Percentage of expression presence/absence variant or its closest neighboring sequence variant that show a maximum linkage disequilibrium estimate r^2_{max} to the SV 100 kb up and downstreams of it (linked) or outside that interval (unlinked) for five r^2 classes.

Therefore, we examined the prediction accuracy that can be obtained when predicting the traits leaf angle, heading date, and plant height, for which h^2 values between 0.69 and 0.76 were observed. In order to obtain unbiased estimates of the prediction accuracy, we randomly subdivided in 1000 cross-validation runs the 23 inbreds in training and validation set. Prediction accuracies of SV, T, and ePAV were compared to the prediction accuracy of the SNParray data set that we used as the baseline predictor. The median prediction accuracy across 1000 cross-validation runs observed for the SNParray data set ranged from -0.49 for heading date to 0.70 for leaf angle (Additional file 1: Figure S4). We observed across the three traits a slightly higher prediction accuracy for the SV extracted from the mRNA sequencing data set compared to the SNParray. An even higher prediction accuracy was observed when using ePAV as predictor. The seedling transcriptome (T_s) resulted across the three traits in the highest median of prediction accuracy of all the examined single predictors.

We also evaluated the pairwise combinations of single predictors and observed for all traits an increase of the prediction accuracy compared to using T_s. Therefore, a grid search in which the relative weights of the relationship matrices of two or three predictors varied in increments of 0.1 prior to summing them up, was used to identify those combinations of SV, ePAV, and T_s that resulted in the highest prediction accuracies. For all three traits, the highest median of the prediction accuracy was observed when using more than one predictor (Fig. 4). Furthermore, a common trend was that the optimal weight of T_s, i.e. the weight that maximizes the prediction accuracy, was at least 40%, whereas the optimal weight of ePAV and SV differed among traits. We examined the prediction accuracy of single predictors as well as optimal combinations of predictors determined from seedling samples sequenced at different depths. Across all traits, we observed that the prediction accuracies decreased for decreasing sequencing depth (Fig. 5). However, the extent of reduction differed between the different predictors and was most pronounced for the SV. The prediction accuracies observed for the optimal combinations of predictors reduced for the three traits only slightly with decreasing sequencing depth. Even with a sequencing depth that corresponds to 0.5% of that of our study, prediction accuracies higher than that of the prediction with the SNParray data set were obtained. However, the variability of the prediction accuracy across the different runs of the resampling simulations increases with a reduced sequencing depth.

Discussion

Transcriptomic variation in barley

Across the 23 inbreds of our study, we have identi-

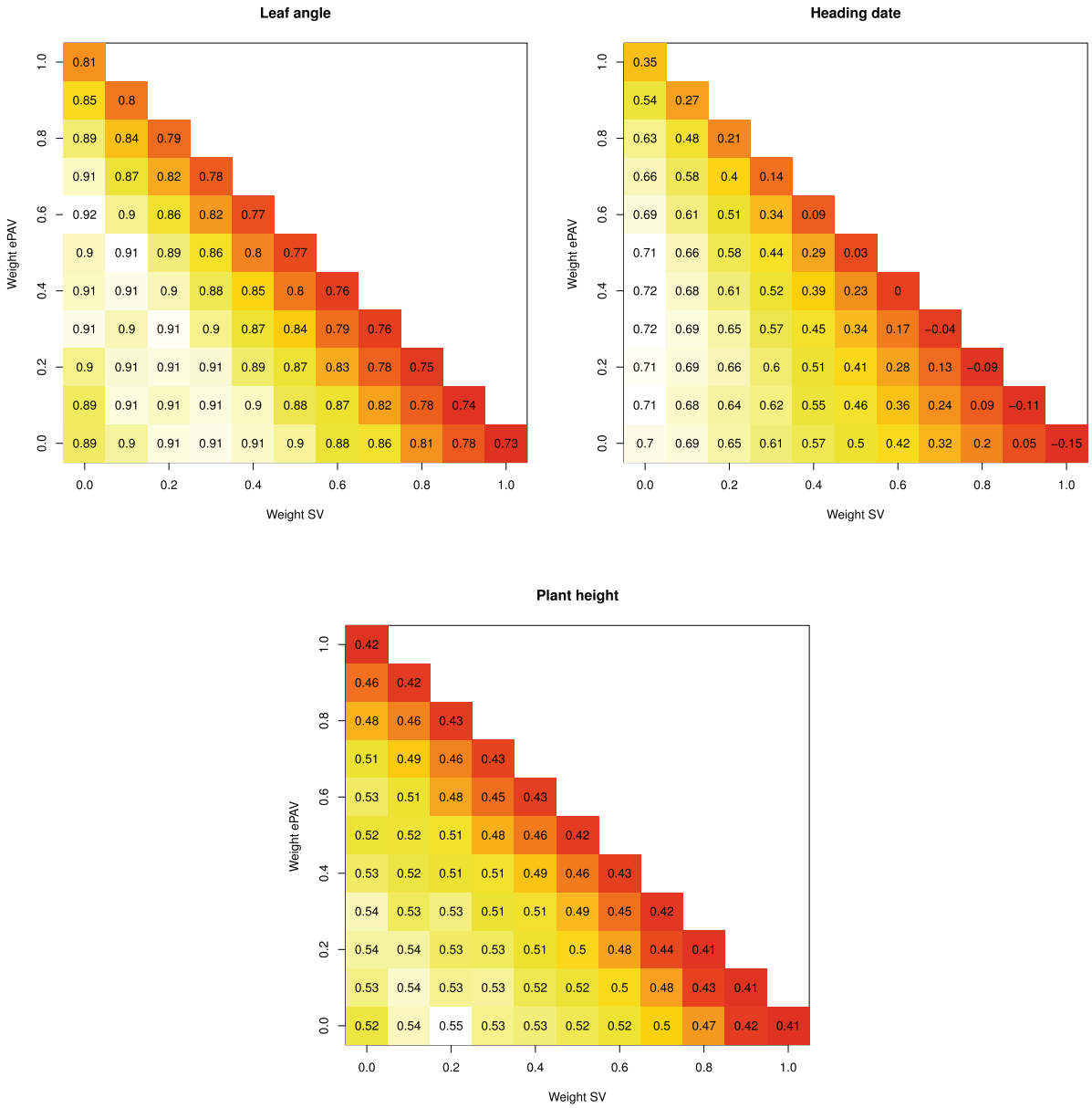

Fig. 4 Prediction accuracy for three different traits. Prediction accuracy for the barley inbreds for leaf angle, heading date, and plant height, for 66 cases which differ in their weights for the predictors sequence variants (SV), expression presence/absence variation (ePAV), and gene expression in seedlings (T_5). Their corresponding relationship matrices were joined with weights varying from 0 to 1 in increments of 0.1. Weights for SV and ePAV are shown at the respective scales; weights for gene expression are = 1 - weight of SV - weight of ePAV. Plotted values represent medians of prediction accuracy across 1,000 cross-validation runs. Heat color schemes differ for the three traits ranging from white, indicating the respective highest value, to red for the respective lowest value

fied 11,523 previously unannotated genes that are not expressed in Morex. Furthermore, we assembled 1,502 newly identified genes that are physically absent from the Morex genome, and are therefore part of barley's dispensable genome. Both numbers are in the range of what was previously reported for barley [17, 18] as well as maize [20, 21]. These genes were added to the standard International Barley Sequencing Consortium (IBSC) gene list and the resulting list was the basis for all following analyses.

Across the three tissues, we observed that about 53% of the total number of genes were detected as ePAV (Additional file 1: Table S2). Despite our lower sample size and the use of three tissues, which both reduce

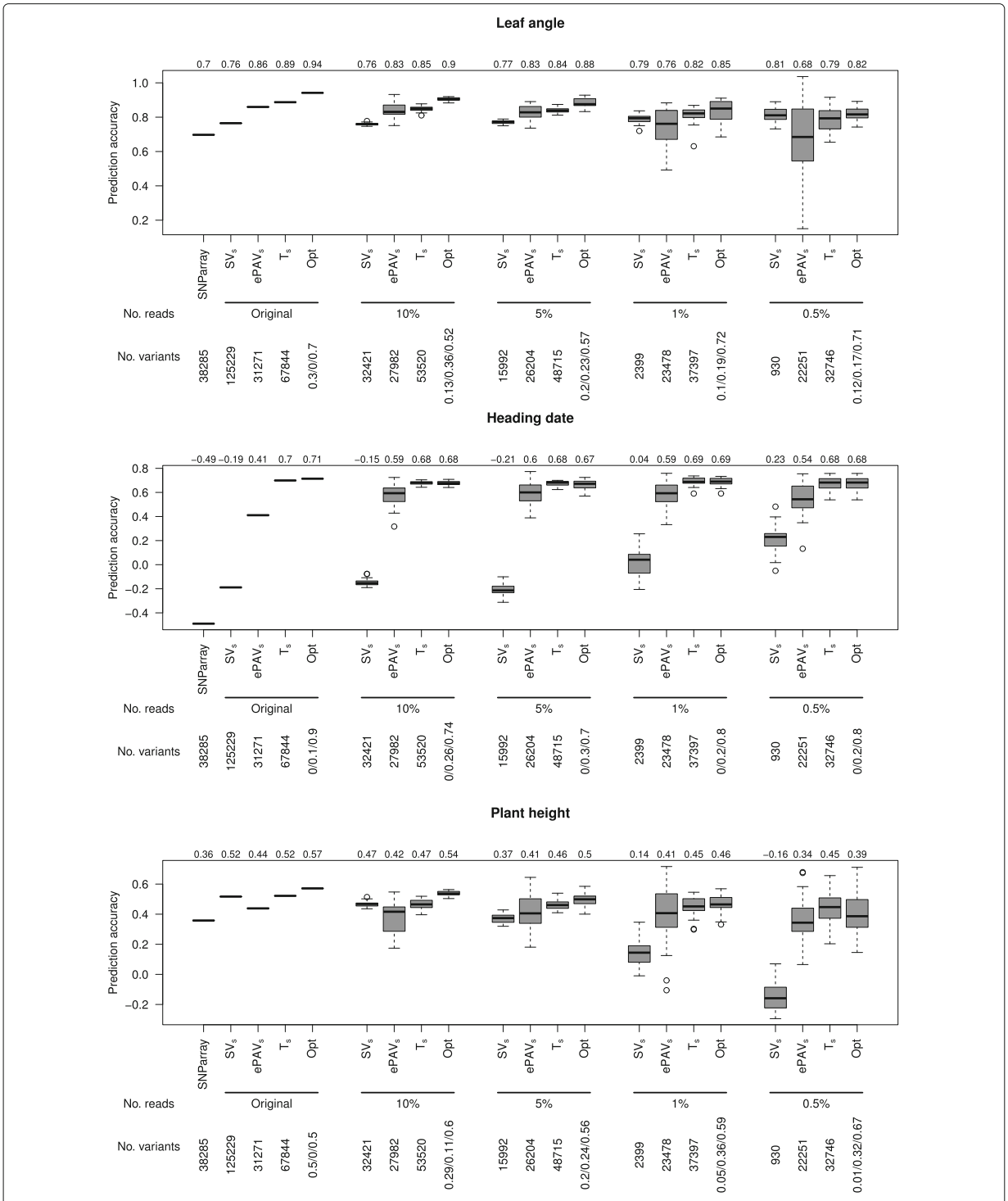

Fig. 5 Prediction accuracy for reduced number of mRNA sequencing reads. Prediction accuracy for the barley inbreds for leaf angle, heading date, and plant height of single predictors as well as the optimal combination identified in a grid search (Opt) using the original number of reads of the seedling sample as well as using data sets for which the number of reads was randomly reduced to 10, 5, 1, and 0.5% of the original number of reads per seedling sample. The number of variants gives the mean number of features available for predictions in each scenario or for the combined predictors the weight of sequence variants (SV_s)/expression presence/absence variation ($ePAV_s$)/gene expression (T_s) resulting in the highest prediction accuracy. The median prediction accuracy is given above each column

the number of detected ePAV, this figure is considerably higher than the maximum 30% that were observed for maize [20, 21]. Because the proportion of ePAV detected here is consistent with other studies in barley [18], the discrepancy of our results and those from maize possibly imply that the proportion of ePAV within a pan-transcriptome is species specific. Due to the unique domestication histories of every crop, selective pressures may have acted differently on dispensable genomes and transcriptomes, especially in cases where PAV variation provided benefits [30]. It is also possible that large genomes rich in repetitive sequences and transposable elements such as the genome of barley contain higher numbers of non-essential genes, whose loss of function has no major impact on plant physiology and can be tolerated by the organism.

We observed that ePAV genes were significantly shorter than an average barley gene. A similar observation was made by Bush et al. [33] in *Arabidopsis thaliana*. Tan et al. [34] described the same trend, although in this case the authors reported variations in gene size (200 bp) that were much smaller than the variations between PAV and non-PAV genes detected by us (5 kb) and by others (1.5 kb) [33]. Another feature that we observed for ePAV genes is that the likelihood that a gene is an ePAV is significantly ($P < 0.05$) unequally distributed across the genome. We observed the highest proportion of ePAV among the present genes in centromeric regions (Fig. 3). This might be explained thereby that selection is less efficient in lowly recombining regions of the chromosome compared to highly recombining pericentromeric regions to purge presence/absence variation that was created by evolutionary processes during plant polyploidization and speciation [30]. gPAV and ePAV were shown to be enriched in genes implicated in disease resistance and stress responses [18, 26, 34–36]. However, the gene ontology (GO) term analysis of the ePAV detected here did not reveal an enrichment of genes implicated in these processes, neither in any other process that could be related to crop performance or adaptation (Additional file 1: Table S3). Further research is required to understand the reason for this difference.

Detecting gPAV by mRNA sequencing

The absence of a gene in a genotype, i.e. the gene is an ePAV, has two possible causes: either the corresponding gene is transcriptionally inactive or it is physically absent from the genome, i.e. it is a gPAV. We were interested in estimating the proportion of ePAV that are due to gPAV. In order to do so, we detected in analogy to Gabur et al. [37] gPAV from the segregation of missing data in biparental populations. This allowed us to estimate that by characterizing the expression of genes in one tissue, we are able to detect about 30% of the existing gPAV. However, a SNP for which a systematic segregation of missing data was observed and that was designated as a gPAV does not necessarily mean that the entire gene in which the SNP is located is missing and therefore not expressed. Instead, the gPAV can be also caused by partial insertion/deletions of the corresponding gene. Therefore, we also examined the power to detect gPAV ($1\text{-}\beta^*$) for a scenario in which only the gene expression in a window of 10 bp around the SNP was considered. In this case of using a single tissue to detect gPAV, the proportion of gPAV that are detected by our ePAV procedure increases to about 46% (Table 1). These findings indicated that our ePAV detection procedure is therefore powerful in detecting gPAV. Furthermore, we observed that $1\text{-}\beta^*$ can be increased even more, if multiple tissues were studied. However, this increase was not of such a size that it justifies the additional resources.

In addition to estimating the power of gPAV detection $1\text{-}\beta^*$, we were also interested in estimating the proportion of ePAV that are not due to gPAV α^*. α^* was approximately 90% in our data set (Table 1), meaning that 10% of the ePAV are caused by the physical absence of the gene and not by impairment of its transcription. This proportion is considerably higher than the 1% reported in maize [21]. An explanation for this finding might be that deletions of entire genes and perhaps of even larger segments of DNA are better tolerated in barley than maize. There is not enough available information on structural variation in the barley genome to be able to compare deletion sizes and frequencies between barley and maize, but it is possible that the presence of long stretches of repetitive elements in barley may have an influence on this process. Another explanation could be the differences in the methodologies between both studies. Jin et al. [21] had used resequencing data to detect gPAV, where our procedure based on patterns of missing data in segregating populations could be more sensitive.

Number of dispensable genes in the barley genome

We can estimate from the above described estimates of $1\text{-}\beta^*$ and α^* that about 10% of the about 38,000 ePAV, i.e. 3,800, are gPAV. With a power $1\text{-}\beta^*$ of about 50% of our ePAV procedure to detect gPAV, the total number of gPAV is expected to be around 7,600 for barley. Therefore, our results suggest that more than 10% of the barley genes show PAV on a genomic level. This figure is similar to what was observed in the analysis of 80 *Arabidopsis* accessions (9%) [34], but was higher compared to other cereal species. Springer et al. [23] estimated that 8.6% of all genes were gPAV in maize. However, their set of analyzed genotypes included, in addition to 19 maize inbreds, 14 wild ancestors, and therefore encompassed a higher genetic diversity compared to our study. As this increases the proportion of detected gPAV, it suggests that cultivated

maize might have a lower gPAV diversity than barley. Consistently, a study in rice also including cultivars and wild relatives estimated to 10% the proportion of gPAV [35], suggesting that rice may also have a lower gPAV diversity than barley. But beyond the proportion of gPAV that a species may contain comes the question of what impact on plant physiology and performance this variation might have. Despite clear examples of gPAV controlling agronomic traits in different species, it has been proposed that most gPAV are not essential and are recent additions to plant genomes [33]. Further data will be necessary to elucidate why enrichment of functional categories occur in certain data sets and not in others, and whether PAV variation has played a more important role in the evolution of crops compared to non-crop species as studies so far seem to suggest [30, 33].

Genomic and transcriptomic prediction

Genomic prediction is becoming a standard tool for plant breeders to increase the gain of selection [38]. The current implementation of genomic selection is mainly based on the use of SNP markers assessed by SNP arrays or genotyping by sequencing methods (for review see Crossa et al. [39]). However, we have observed considerable variation of T as well as ePAV and, very importantly, this variation was largely independent from the variation explained by neighboring SNP (Table 2). Therefore, the accuracy of T and ePAV to predict phenotypic traits was assessed.

We have observed that for all three examined traits both types of information that were extracted from the mRNA sequencing data set, the SV as well as the ePAV, resulted in higher prediction accuracy when using GBLUP compared to the classically used SNP data generated with a 50K SNParray (Additional file 1: Figure S4). For SV, that might be explained by the higher number of features compared to the SNParray information, which in turn increases the extent of LD between the SNP and the QTL [40]. However, for the ePAV this was not the case. Instead, the superiority of the ePAV information compared to the SNParray for the prediction of phenotypic traits might be due to that ePAV are only caused to 10% by gPAV but cover also gene expression differences. The transcriptome T is expected to incorporate gene expression and physiological epistasis [41] and therefore has a considerably higher prediction accuracy compared to SV or SNParray (Additional file 1: Figure S4), even when modelling statistical epistasis.

However, we also observed differences in the prediction accuracy of T depending on the tissue that was used for mRNA extraction. The prediction accuracies were on average across the three examined traits considerably higher for T_s compared to T_l data set. This finding might be explained either by the fact that the number of cell types that were included for mRNA extraction were more diverse for the former than the latter and thereby increases the number of features from 60,888 to 67,844. Another non mutually exclusive possible explanation is that the time of heterogenous environmental factors to influence the genotypes was lower for the seedling samples compared to the leaf samples. And in the set-up used in our study of unreplicated plants for sample collection such heterogenous environmental factors cause together with genotype*environment interaction a reduction of the precision of the measurement of the predictor. This in turn is expected to reduce the prediction accuracy. Our finding indicated that the transcriptome of seedlings grown on filter paper is a good proxy of the gene activity for a broad range of developmental stage of plants grown in a diverse set of environments.

Schrag et al. [29] derived from a comparison of pairs of single predictors with their combinations the following two conclusions. First, combining the best single predictor for a certain trait with another predictor did not improve predictions and in some cases rather impaired predictive ability. Second, combinations that did not comprise the best single predictor tended to be superior to both components individually. Both of them were not in agreement with our findings. Instead, we have observed a complementarity between the best single predictor T_s and SV and even between T_s and ePAV (Additional file 1: Figure S4).

Therefore, a grid search was used to identify those combinations of SV, ePAV, and T_s that maximize the prediction accuracy. For all three traits, the highest median of the prediction accuracy was observed when using more than one predictor (Fig. 4). In contrast to the results of Schrag et al. [29] and Xu et al. [35], we have observed rather small differences between the optimal weight of the three predictors across the three examined traits, despite that these were assessed at completely different developmental stages. The likely explanation for this difference is that, in contrast to Schrag et al. [29] and Xu et al. [35], we focused on genetic and transcriptional predictors and did not include features derived from metabolome analyses, which represent a completely different level of information.

Application in breeding

In the above described grid search, the SV and ePAV data sets were extracted from the mRNA sequencing data of multiple tissues. A cost efficient integration of our approach in practical breeding programs would require that all data sets are extracted from the sequence experiment of one tissue. Due to the above described quantitative genetic advantage of the seedling sample but also the logistical advantages of using seedling samples that are generated on filter paper in petri dishes: they require a much lower amount of space, personnel and material resources, allow a season independent cultivation, as well as can be generated faster

as the turn over time is shorter, they were studied in detail.

The prediction accuracy of the original sequencing depth was not influenced by predicting the phenotypic traits from SV and ePAV features extracted from the seedling sample instead from the three tissues. This can be explained by the fact that the SV_s can be adequately assessed also with one tissue and differences between $ePAV_s$ and ePAV are compensated for by T_s. The prediction accuracy observed for this scenario is considerably higher compared to using the SNParray information. However, the cost of genotyping one sample with the barley 50K SNParray is with about 50 Euro [42] also less than the mRNA sequencing analysis. When generating it newly with latest protocols and sequencing chemistry one could expect that the mRNA sequencing of one sample would cost about 2 Euro for the mRNA library preparation [43] as well as 60 Euro for 20 million 2x150 bp reads. In addition, breeding companies use for their routine genotyping in many cases smaller SNP arrays than the one used in our study. This would reduce the costs even more, but will decrease the prediction accuracy, especially, if diverse genetic material is used [44] as in our study. Therefore, we performed downsampling simulations to examine the reduction of the prediction accuracy if the sequencing depth is reduced.

The prediction accuracies observed for the optimal combinations of single predictors reduced for the three traits only slightly with a decreasing sequencing depth. The main limitation to reducing the sequencing depth to values below 1% of that of our study, i.e. about $2x10^5$ 2x150 bp reads, is not the reduction of the median of the prediction accuracy but the increasing standard deviation (Fig. 5). This increase is caused by the increasing sampling variance of the low depth sequencing. However, our results indicate that down to 5% of our data set, i.e. about $1x10^6$ 2x150 bp reads, the obtained prediction accuracy was in more than 95% of the resampling runs higher than that obtained with the SNParray data set. Such a transcriptomic characterization would cost about 5 Euro and is therefore also less expensive than current GBS approaches with the advantage of higher prediction accuracies. Therefore, we consider mRNA sequencing based characterizations of breeding material harvested as seedlings in petri dishes as a powerful and cost efficient approach to replace current SNP based characterization. For species that are bred in breeding categories other than inbred lines, the phenotypic evaluation is even more expensive [45] than for species bred as inbred lines. Therefore, an approach as suggested above will increase the gain of selection for such species even more, as the cost advantage is higher than in species bred as inbred lines.

Conclusion

We have used mRNA sequencing as an approach to explore the dispensable genome and transcriptome of barley in 23 spring barley inbreds, and estimate that 53% of genes are ePAV. Our analyses suggest that about 10% of ePAV in barley are due to the physical absence of a gene in an inbred (gPAV). We have observed that the omic variation that was extracted from the mRNA sequencing data set, the sequence variants (SV), the ePAV, as well as the transcriptome (T) resulted individually in higher prediction accuracies compared to the classically used SNParray data set. This superiority was even more pronounced when using optimal combinations of SV, ePAV, and T to predict phenotypic traits. Finally our results suggest that low coverage mRNA sequencing based characterization of breeding material harvested as seedlings in petri dishes is a powerful and cost efficient approach to replace current SNP based characterization.

Methods

Plant material

Our analyses were based 23 spring barley inbreds that were selected out of a worldwide collection of 224 inbreds [31] (Additional file 1: Table S1) using the MSTRAT algorithm [46]. For these inbreds, the maximal combined genotypic and phenotypic richness index was observed. Seeds of the 23 spring barley inbreds were sown in controlled greenhouse conditions with 16 hours light and eight hours dark at 22 °C. A fragment of the youngest fully developed leaf from two different plants was collected for each inbred. The collection of all samples was done within one hour to minimize the variation due to circadian rhythms. For a total of six inbreds, apices were harvested at stage 47 of the Zadoks scale [47]. Young seedlings were harvested in an independent experiment. Seeds were surface sterilized with 1% bleach and rinsed with sterile water. Eight seeds per inbred were placed between two layers of sterile filter paper soaked with 5 mL of sterile water. The petri dishes were placed in the greenhouse with the above described environmental conditions. Five days after germination, two seedlings were sampled for each inbred. All collected samples were immediately flash frozen in liquid nitrogen. The above described experiments were performed in accordance to the experimental design of related studies [20, 21] with one biological replicate only, as the replication of alleles is provided among genotypes.

For the assessment of phenotypic traits under field conditions, the 23 spring barley inbreds were planted as replicated check genotypes in an experiment with other entries which was layed out as an augmented row column design. This experiment was performed in three environments (Cologne 2017 and 2018 and Mechernich 2018) as single row plots with 10 plants/plot as well as in a fourth environment (Quedlinburg 2018) as

double row plots with 40 plants/plot. At each of the four agro-ecologically diverse environments in Germany, the 23 barley inbreds were replicated 21, 20, 19, and 19 times, respectively. For each experimental plot, three traits were assessed. The leaf angle of about four weeks old plants was scored on a scale from 1 to 9, where 1 indicates erect leaves and 9 prostrate leaves. The heading date was assessed as number of days after planting. Furthermore, the plant height in cm was assessed after heading.

SNP genotyping and quantification of gene expression

An Illumina 50K barley SNP array [5] was used to genotype the 23 inbreds. This data set is designated in the following as SNParray. The same array was used to genotype between 35 and 146 F5 progenies of 45 segregation populations which were derived from double-chain crosses [32] of the 23 inbreds (Casale et al. in preparation).

mRNA was extracted from leaf, seedling, and apex samples (cf. Digel et al. [16]). A total of 52 polyA enriched mRNA libraries were prepared. The 150 bp paired-end Illumina sequencing libraries with individually barcoded samples were sequenced on an Illumina HiSeq2000 sequencer (Illumina, Inc., San Diego, CA USA). Reads were trimmed using trim_galore and then mapped against the unmasked barley reference sequence [6] using HISAT2 [48]. Trinity was used to perform a *de novo* assembly of the unmapped reads of all inbreds [49]. The assembled contigs that were expressed in at least two tissue samples were BLASTn-searched against a human and viral database, to exclude contigs that are due to contaminations (e-value $\leq$1e-5, identity $\geq$95.0%). Then, the contigs were searched against a barley database to remove, based on the same thresholds, genes which are too similar compared to barley reference genes. All contigs that had a homology (e-value $\leq$1e-5, identity $\geq$98.0%) to an annotated protein in at least one of the species *Arabidopsis thaliana*, *Brachypodium distachyon*, *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, *Triticum aevisticum*, *Triticum dicoccum*, and *Secale cereale* were retained. The contig with the maximum coverage was chosen as representative contig for each gene [6]. These contigs were designated as newly identified genes.

Transcript calling was performed with StringTie [50] using a gene annotation file that comprised low and high confidence genes of transcripts defined in the barley reference genome [6] and the newly identified genes of the *de novo* assembly.

Genes which mapped to the reference sequence and were expressed in at least two samples, but which were not available in the IBSC-reference annotation file were designated in the following as newly annotated genes. The gene expression quantified as fragments per kilobase of exon model per million fragments mapped (FPKM) is designated in the following as T, where the indexes *l*, *s*, *a* were used to separate the tissues leaf, seedling, and apex.

Identification of ePAV

For each tissue, a presence call was made for each inbred-gene combination in the matrix of presence/absence calls, if T >0 and an absence call if T = 0. No call was made for the inbreds with 0< T <10% of the maximum value of T for a gene-tissue combination (cf. Jin et al. [21]). Tissue specific ePAV calls were combined to an across tissue ePAV call as follows: If the presence/absence call made for all tissues of one inbred-gene combination was identical, this call was kept. For all inbred-gene combinations with a presence call for at least one tissue, a presence call was kept in the across tissue matrix of presence/absence calls. An absent call was kept in the across tissue matrix of presence/absence calls for all inbred-gene combinations with only no or absent calls across tissues. These genes were designated in the following as ePAV which have an across tissue ePAV call of present and absent each for at least two inbreds (cf. Jin et al. [21]).

We used in analogy to Gabur et al. [30] the segregation of missing data in biparental populations to determine the percentage of ePAV that are due to gPAV. For the SNP from the SNParray dataset for which no missing data was observed, the Q_{90} of the major allele frequency was calculated per population to consider random deviations from an allele frequency of 0.5. For each population, each SNP was assigned to one of three categories based on the proportion of missing data: A: [0, Q_{90}), B: [Q_{90},1-Q_{90}], C: (1-Q_{90}, 1]. Category A to C can be interpreted as both parental inbreds have a present call, one parental inbred has a present and one an absent call, both parental inbreds have an absent call, respectively. A parental inbred was assigned an absent call at a SNP, if all populations derived from that parent were of category B or C. A parental inbred was assigned a present call at a SNP, if all populations derived from that parent were of category A or B. These 1,972 SNP that have a present and absent each for at least one inbred were designated in the following as gPAV-SNP (Additional file 1: Figure S2). A total of 14,843 barley genes comprised in their coding sequence one SNP from the SNParray and were designated in the following as genic SNP.

The 1,105 gPAV-SNP that were genic SNP and that were not within 30 bp of an insertion were designated as genic PAV-SNP.

The statistical power (1-β^*) to detect gPAV by mRNA sequencing was calculated as the percentage of genic PAV-SNP that were located within the coding sequence of ePAV. Furthermore, the empirical type I error (α^*) of our ePAV procedure was estimated as the proportion of genes that comprised a genic SNP, no genic PAV-SNP, but

were detected as ePAV out of the total number of detected ePAV. In addition, we calculated the proportion of correct allele assignments (o) as the proportion of common presence/absence ePAV calls and presence/absence calls at genic PAV-SNP across all 23 inbreds. We estimated 1-β^* and α^* firstly for ePAV determined based on T of the entire gene as well as based on T calculated for 10 bp large windows surrounding the genic SNP.

A resampling procedure was used to determine the robustness of our ePAV detection procedure. For each gene, randomly 20% of the entire gene length were selected and transcript calling and ePAV detection were performed. This was repeated 50 times and the average of number and proportion of detected ePAV was calculated. The null hypothesis of a uniform distribution of ePAV across the genome and chromosomes was tested by a permutation procedure. The difference of mean gene length of ePAV and non-ePAV was tested for its statistical significance using a t-test. GO term enrichment analysis of ePAV was performed using the R-package topGO [51]. GO terms of newly annotated genes and newly identified genes were defined based on those available from the agriGO-database of homologous genes as described by Mascher et al. [6].

Population genetic analyses

Variant calling was performed with samtools and bcftools. Sequence variants "SV" with mapping quality < 55 were removed from the analysis. If the sequencing depth of a SV was smaller than 5, the allele call was set to "NA". SV with a heterozygosity >10% were discarded and the alleles at remaining heterozygous sequence variants were set to "NA" for the corresponding inbreds. Biallelic sequence variants with a maximum of 20% missing information were retained. If the allele call was different between the tissues of the same inbred, the call with the higher sequencing depth was retained. Missing values in the matrices of SV and ePAV were mean imputed.

Associations among inbreds based on SV, ePAV, as well as T were revealed with a principal component analysis [52]. Pearson's correlation coefficients were calculated between euclidean distance matrices of SV, ePAV, and T. Linkage disequilibrium measured as r^2 [53] was calculated between ePAV and linked/unlinked SV.

Genomic prediction

Each of the three phenotypic traits leaf angle, heading date, and plant height was analyzed across the four environments using mixed models. This allowed to estimate adjusted entry means as well as the heritability on an entry mean basis.

The adjusted entry mean of each barley inbred for each trait was predicted using genomic best linear unbiased prediction (GBLUP) [54–57]. GBLUP was used as implemented in the R-package sommer [58], where only additive effects were modeled and the residuals were assumed to be normally distributed with mean 0 and variance σ_e^2.

The performance of the barley inbreds was predicted using different predictors: (i) SNParray, (ii) SV, (iii) ePAV, (iv) T_l, (v) T_s. W is a matrix of feature measurements for the respective predictors. The dimension of W is determined by the number of barley inbreds and the number of features in the corresponding predictor ($m_{SNParray}$ = 44,045 m_{SV} = 133,566, m_{ePAV} = 38,810, m_{Tl} = 60,888, m_{Ts} = 67,844). The columns in W were centered and standardized to unit variance. For each predictor, an additive relationship matrix G was calculated according to VanRaden [59]. The matrices G of two or three predictors were weighted and summed up, resulting in one joined weighted relationship matrix [29]. A grid search, varying the relative weights in increments of 0.1, resulted in 66 different joined weighted relationship matrices. We calculated the prediction accuracy $[r(\hat{g},g)]$ for each examined scenario.

The standard scheme for validation of genomic prediction was five-fold cross-validation. For this purpose, the 23 inbreds were randomly subdivided into five disjoint subsets. One subset was left out for validation, whereas the other four subsets were used as training set. This procedure was replicated 200 times, yielding a total of 1000 cross-validation runs. The median of the prediction accuracy across the 1,000 cross-validation runs was calculated. From the original data set of seedling samples, the number of reads was randomly reduced to 10, 5, 1, and 0.5% of the original number of reads per inbred. This procedure was replicated 30 times. For these subsets of reads, the above described work flow of read mapping, determination of gene expression, expression presence/absence variation, and sequence variant calling was performed. The prediction accuracy for the single predictors SV_s, $ePAV_s$, and T_s and the combination of these predictors, was calculated for leaf angle, heading date, and plant height as average across the 30 replications.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6174-3.

Additional file 1: Supplementary Table S1: Summary of barley inbreds. **Supplementary Table S2:** Number of expression presence/absence variation (ePAV) observed for our detection procedure. **Supplementary Table S3:** Gene ontology term enrichment analysis. **Supplementary Fig. S1:** Characterization of the not annotated contigs established by the transcript calling. **Supplementary Fig. S2:** Procedure to evaluate the detection of presence/absence variation. **Supplementary Fig. S3:** Population structure of the 23 barley inbreds. **Supplementary Fig. S4:** Prediction accuracy of single predictors.

Abbreviations
ePAV: Expression presence/absence variation; FPKM: Fragments per kilobase of exon model per million fragments mapped; GBLUP: Genomic best linear unbiased prediction; GO: Gene ontology; gPAV: Genomic presence/absence variation; IBSC: International Barley Sequencing Consortium; LD: Linkage disequilibrium; PAV: Presence/absence variation; PCA: Principal component analysis; SNP: Single nucleotide polymorphism; SV: Sequence variant; T: Transcriptomic variation; T_a: Apex transcriptomic variation; T_l: Leaf transcriptomic variation; T_s: Seedling transcriptomic variation

Acknowledgements
The authors thank the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) for creating and sequencing the mRNA libraries of this study. Furthermore, we would like to thank Florian Esser, Marianne Harperscheidt, and Anja Kyriacidis for technical assistance with performing the field experiments at Cologne and Mechernich as well as Dr. Franziska Wespel (Saatzucht Breun) and her team for realizing the field experiment at Quedlinburg. We are grateful to Arno Hamacher for providing the experimental field in Mechernich. This paper is dedicated to Professor Dr. Albrecht E. Melchinger on the occasion of his 70th anniversary.

Authors' contributions
BS designed and coordinated the project; AdM collected the materials; AdM prepared and purified mRNA and DNA samples; MW, AdM, MP, DR, and BS performed the data analyses; MW, AdM, and BS wrote the paper. All authors read and approved the final manuscript.

Funding
This research was funded by the the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111).

Availability of data and materials
The sequence data have been deposited in the NCBI Sequence Read Archive (SRA) under accession PRJNA534414. SV (SV_genome.csv), SNParray data (SNParray.csv), phenotypic data (aem_alltraitsDRRobs1718.csv), expression data (T_leaf.csv, T_seed-ling.csv), ePAV data (ePAV_consensus.csv, ePAV_leaf.csv, ePAV_seedling.csv) are contained within the paper and its additional files.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 11 July 2019 Accepted: 6 October 2019
Published online: 29 October 2019

References

1. Beddington RJ, Asaduzzaman M, Clarke ME, Bremauntz AF, Guillou MD, Howlett D, Jahn MM, Lin E, Mamo T, Negra C, Nobre CA, Scholes RJ, Bo NV, Wakhungu J. What next for agriculture after durban? Science. 2012;335:289–90.
2. Zohary D, Hopf M, Weiss E. Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin. Oxford: Oxford Univ Press; 2012.
3. Dawson IK, Russell J, Powell W, Steffenson B, Thomas WTB, Waugh R. Barley: A translational model for adaptation to climate change. New Phytol. 2015;206:913–31.
4. Nevo E, Fu Y-B, Pavlicek T, Khalifa S, Tavasi M, Beiles A. Evolution of wild cereals during 28 years of global warming in Israel. Proc Nat Acad Sci. 2012;109:3412–5.
5. Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, Russell J, Shaw PD, Thomas W, Waugh R. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 2017;8:1792.
6. 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 X-q, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544:427–33.
7. Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N. Genebank genomics highlights the diversity of a global barley collection. Nature Genet. 2019;51:319–26.
8. Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP. A new resource for cereal genomics: 22K barley GeneChip comes of age. Plant Physiol. 2004;134:960–8.
9. Druka A, Muehlbauer G, Druka I, Caldo R, Baumann U, Rostoks N, Schreiber A, Wise R, Close T, Kleinhofs A, Graner A, Schulman A, Langridge P, Sato K, Hayes P, McNicol J, Marshall D, Waugh R. An atlas of gene expression from seed to seed through barley development. Function Integr Genomics. 2006;6:202–11.
10. Druka A, Potokina E, Luo Z, Bonar N, Druka I, Zhang L, Marshall DF, Steffenson BJ, Close TJ, Wise RP, Kleinhofs A, Williams RW, Kearsey MJ, Waugh R. Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* in barley. Theoret Appl Genet. 2008;117:261–272.
11. Chen X, Hackett CA, Niks RE, Hedley PE, Booth C, Druka A, Marcel TC, Vels A, Bayer M, Milne I, Morris J, Ramsay L, Marshall D, Cardle L, Waugh R. An eQTL analysis of partial resistance to *Puccinia hordei* in barley. PLoS ONE. 2010;5:8598.
12. Greenup AG, Sasani S, Oliver SN, Walford SA, Millar AA, Trevaskis B. Transcriptome analysis of the vernalization response in barley (*Hordeum vulgare*) seedlings. PLoS ONE. 2011;6:17900.
13. Hemming MN, Walford SA, Fieg S, Dennis ES, Trevaskis B. Identification of high-temperature-responsive genes in cereals. Plant Physiol. 2012;158:1439–50.
14. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J. 2008;53:90–101.
15. Potokina E, Druka A, Luo Z, Moscou M, Wise R, Waugh R, Kearsey M. Tissue-dependent limited pleiotropy affects gene expression in barley. Plant J. 2008;56:287–96.
16. Digel B, Pankin A, von Korff M. Global transcriptome profiling of developing leaf and shoot apices reveals distinct genetic and environmental control of floral transition and inflorescence development in barley. Plant Cell. 2015;27:2318–34.
17. Takahagi K, Uehara-Yamaguchi Y, Yoshida T, Sakurai T, Shinozaki K, Mochida K, Saisho D. Analysis of single nucleotide polymorphisms based on RNA sequencing data of diverse bio-geographical accessions in barley. Nature Sci Rep. 2016;6:33199.
18. Ma Y, Liu M, Stiller J, Liu C. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. BMC Genomics. 2019;20:12.
19. Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J. Genome-wide patterns of genetic variation among elite maize inbred lines. Nature Genet. 2010;42:1027–30.
20. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza A, Barry K, Leon ND, Kaeppler SM, Buell CR. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014;26:121–35.
21. Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Nature Sci Rep. 2016;6:18936.
22. Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS. All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc Nat Acad Sci. 2006;103:6805–10.
23. Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS. Maize inbreds exhibit high levels of copy

number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 2009;5:1–17.
24. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C, Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nature Genet. 2018;50:1289–95.
25. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y, Fan D, Zhao Y, Wang Z, Zhou C, Chen J, Zhu C, Li W, Weng Q, Xu Q, Wang Z-X, Wei X, Han B, Huang X. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature Genet. 2018;50:278–84.
26. 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. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. 2013;14:58.
27. Riedelsheimer C, Altmann T, Grieder C, Technow F, Stitt M, Lisec J, Riedelsheimer C, Willmitzer L, Sulpice R, Melchinger AE, Czedik-Eysenberg A. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature Genet. 2012;44:217–220.
28. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE. Genomic predictability of interconnected biparental maize populations. Genetics. 2013;194:493–503.
29. Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. Beyond genomic prediction: Combining different types of omics data can improve prediction of hybrid performance in maize. Genetics. 2018;208:1373–85.
30. Gabur I, Chawla HS, Snowdon RJ, Parkin IAP. Connecting genome structural variation with complex traits in crop plants. Theoret Appl Genet. 2019;132:733–50.
31. Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH. Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breed. 2010;129:271–9.
32. Stich B. Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. Genetics. 2009;183:1525–34.
33. Bush SJ, Castillo-Morales A, Tovar-Corona JM, Chen L, Kover PX, Urrutia AO. Presence-absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. Mole Biol Evol. 2013;31:59–69.
34. Tan S, Zhong Y, Hou H, Yang S, Tian D. Variation of presence/absence genes among *Arabidopsis* populations. BMC Evol Biol. 2012;12:86.
35. 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. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnol. 2012;30:105–11.
36. Jiang L, Ge M, Zhao H, Zhang T. Analysis of heterosis and quantitative trait loci for kernel shape related traits using triple testcross population in maize. PLoS ONE. 2015;10:0124779.
37. Gabur I, Chawla HS, Liu X, Kumar V, Faure S, von Tiedemann A, Jestin C, Dryzska E, Volkmann S, Breuer F, Delourme R, Snowdon R, Obermeier C. Finding invisible quantitative trait loci with missing data. Plant Biotechnol J. 2018;16:2102–12.
38. Desta ZA, Ortiz R. Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci. 2014;19:592–601.
39. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. Genomic selection in plant breeding: Methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.
40. Goddard ME, Hayes BJ. Genomic selection. J Animal Breed Genet. 2007;124:323–30.
41. Sackton TB, Hartl DL. Genotypic context and epistasis in individuals and populations. Cell. 2016;166:279–87.
42. Monat C, Schreiber M, Stein N, Mascher M. Prospects of pan-genomics in barley. Theoret Appl Genet. 2019;132:785–96.
43. Alpern D, Gardeux V, Russeil J, Deplancke B. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biol. 2019;20:70.
44. Stich B, Inghelandt DV. Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. Front Plant Sci. 2018;9:159.
45. Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 2010;50:1681–90.
46. Gouesnard B. MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J Heredity. 2001;92:93–4.
47. Zadoks JC, Chang TT, Konzak CF. A decimal code for the growth stages of cereals,. Weed Res. 1974;14:415–21.
48. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12:357–60.
49. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494–512.
50. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnol. 2015;33:290–5.
51. Alexa A, Rahnenfuhrer J. topGO: Enrichment analysis for gene ontology. R package. 2018. R package:version 2.34.0.
52. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53:325–38.
53. Hill WG, Robertson A. Linkage disequilibrium among neutral genes in finite populations. Theoret Appl Genet. 1968;38:226–31.
54. Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics. 2002;161:373–9.
55. Park T, Casella G. The bayesian lasso. J Am Stat Assoc. 2008;103:681–6.
56. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res. 2009;91:47–60.
57. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.
58. Covarrubias-Pazaran G. Genome-Assisted prediction of quantitative traits using the r package sommer. PLoS ONE. 2016;11:0156744.
59. VanRaden PM. Efficient Methods to Compute Genomic Predictions. J Dairy Sci. 2008;91:4414–23.

Publisher's Note

SUPPLEMENTARY INFORMATION

Table S1: Inbred lines included in this study, their country of origin (CoO), row type, year of release, and the sequenced tissues.

					RNA sequencing		
Inbred name	BCC code	CoO	Row type	Year of release	Leaf	Seedling	Apex
HOR1842	HOR1842	AFG	6	1935	x	x	
HOR383	BCC1561	BGR	6	unknown	x	x	
Sanalta	BCC929	CAN	2	1930	[1]	[1]	
ItuNative	BCC502	CHN	6	unknown	x	x	
Sissy	BCC1413	GER	2	1990	x	x	x
Georgie	BCC1381	GBR	2	1975	x	x	
SprattArcher	BCC1415	GBR	2	1943	x	x	x
Lakhan	BCC533	IND	6	unknown	x	x	
Kharsila	HOR11403	IND	6	before 1911	x	x	
W23829/803911	HOR11374	ISR	2	unknown	x	x	x
Namhaebori	BCC667	KOR	6	unknown	x	x	
IG128216	BCC118	LBY	6	1983	x	[1]	
IG128104	BCC173	PAK	6	1974	x	x	
K10693	BCC1491	RUS	6	unknown	x	x	
IG31424	BCC190	SYR	2	1981	x	x	
HOR12830	HOR12830	SYR	6	unknown	x	x	
HOR7985	HOR7985	TUR	2	before 1969	x	x	x
K10877	BCC1503	TKM	6	unknown	x	x	x
HOR8160	HOR8160	TUR	2	before 1969	x	x	
Ancap2	BCC807	URY	6	1950	x	x	
CM67	BCC846	USA	6	1983	x	x	
Kombyne	BCC893	USA	6	1975	[1]	x	
Unumli-Arpa	BCC1470	UZB	2	unknown	x	x	x

[1] Samples were removed during the data cleaning process

Table S2: Number of expression presence/absence variation (ePAV) observed for our detection procedure.

Data set	#ePAV	#Genes	ePAV [%]
Barley, All	38,810	73,187	53.0
Barley, IBSC	28,340	60,162	47.1
Barley, newly annotated	9,286	11,523	80.6
Barley, newly identified	1,184	1,502	78.8

Table S3: The 15 GO terms of biological process that were most significantly enriched for ePAV.

GO.ID	Term	#Significant genes	#Expected genes	p-value
GO:0015074	DNA integration	761	166.95	< 1e-30
GO:0055114	oxidation-reduction process	705	241.69	< 1e-30
GO:0055085	transmembrane transport	251	119.07	< 1e-30
GO:0006278	RNA-dependent DNA biosynthetic process	337	142.98	< 1e-30
GO:0006605	protein targeting	105	48.82	< 1e-30
GO:0006508	proteolysis	630	433.82	< 1e-30
GO:0006468	protein phosphorylation	705	481.82	< 1e-30
GO:0006333	chromatin assembly or disassembly	170	99.18	< 1e-30
GO:0044238	primary metabolic process	4000	3255.6	< 1e-30
GO:0048544	recognition of pollen	77	19.28	1.1e-23
GO:0006281	DNA repair	144	99	1.3e-23
GO:0008152	metabolic process	5080	3967.22	4.9e-20
GO:0006313	transposition, DNA-mediated	46	8.42	7.1e-20
GO:0000723	telomere maintenance	60	18.54	1.6e-16
GO:0016358	dendrite development	20	8.78	1.1e-15

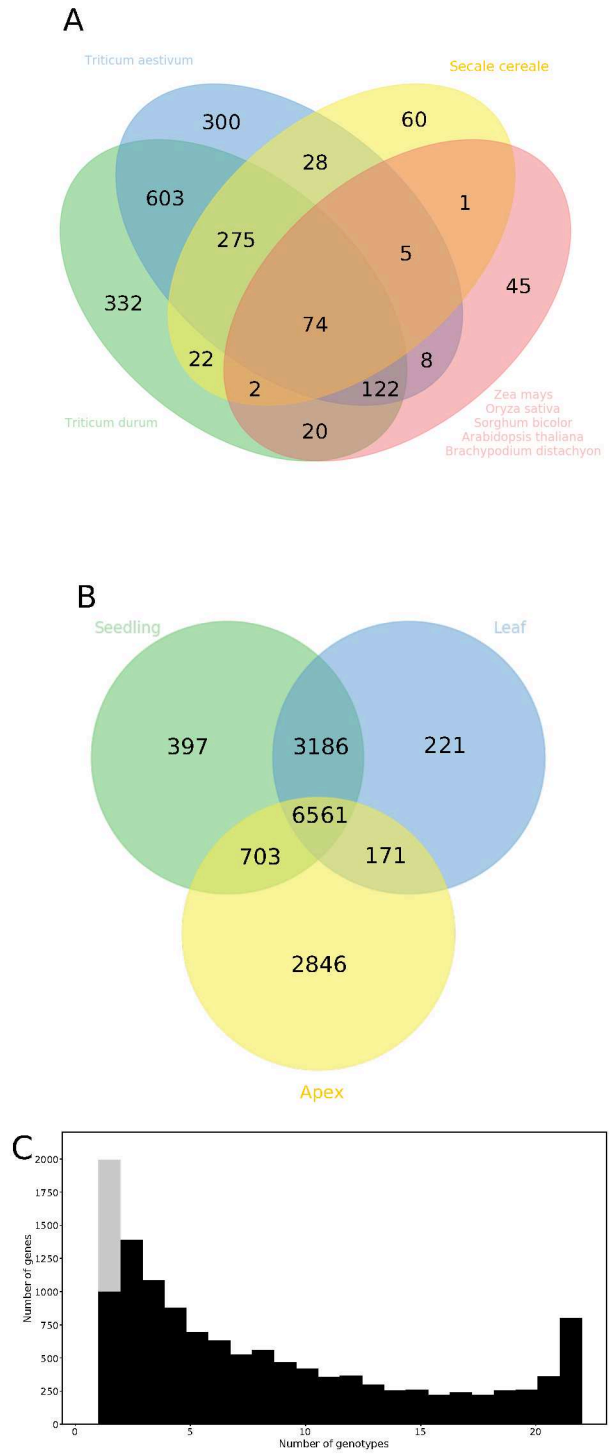

Fig. S1: Characterization of the not annotated contigs established by the transcript calling. A) Newly annotated genes which had based on BLASTn searches homology to eight different plant species (1,897). B) Expression of 11,523 newly annotated genes in the three different tissues. C) Number of inbred lines in which the not annotated contigs were called during the transcript calling. Gray bar shows the contigs detected in only one sample. The 11,523 genes, which were expressed in at least two samples, were marked in black and were designated as new annotated genes.

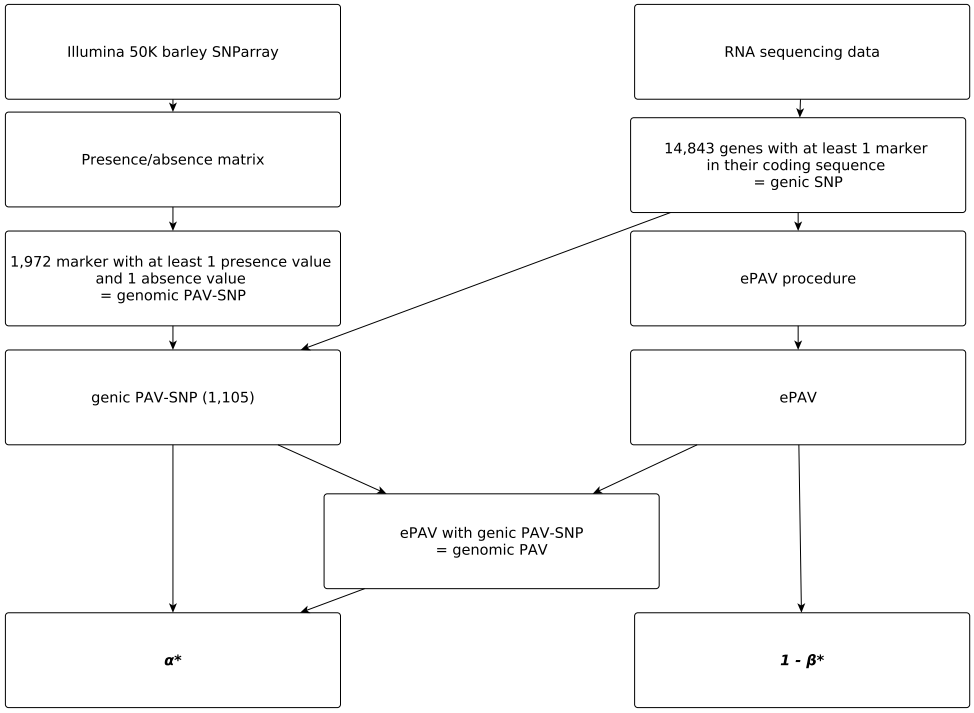

Fig. S2: Overview of the process to estimate the statistical power (1-β^*) and the empirical type I error rate (α^*) to detect genomic presence/absence variation (gPAV) by expression presence/absence variation (ePAV).

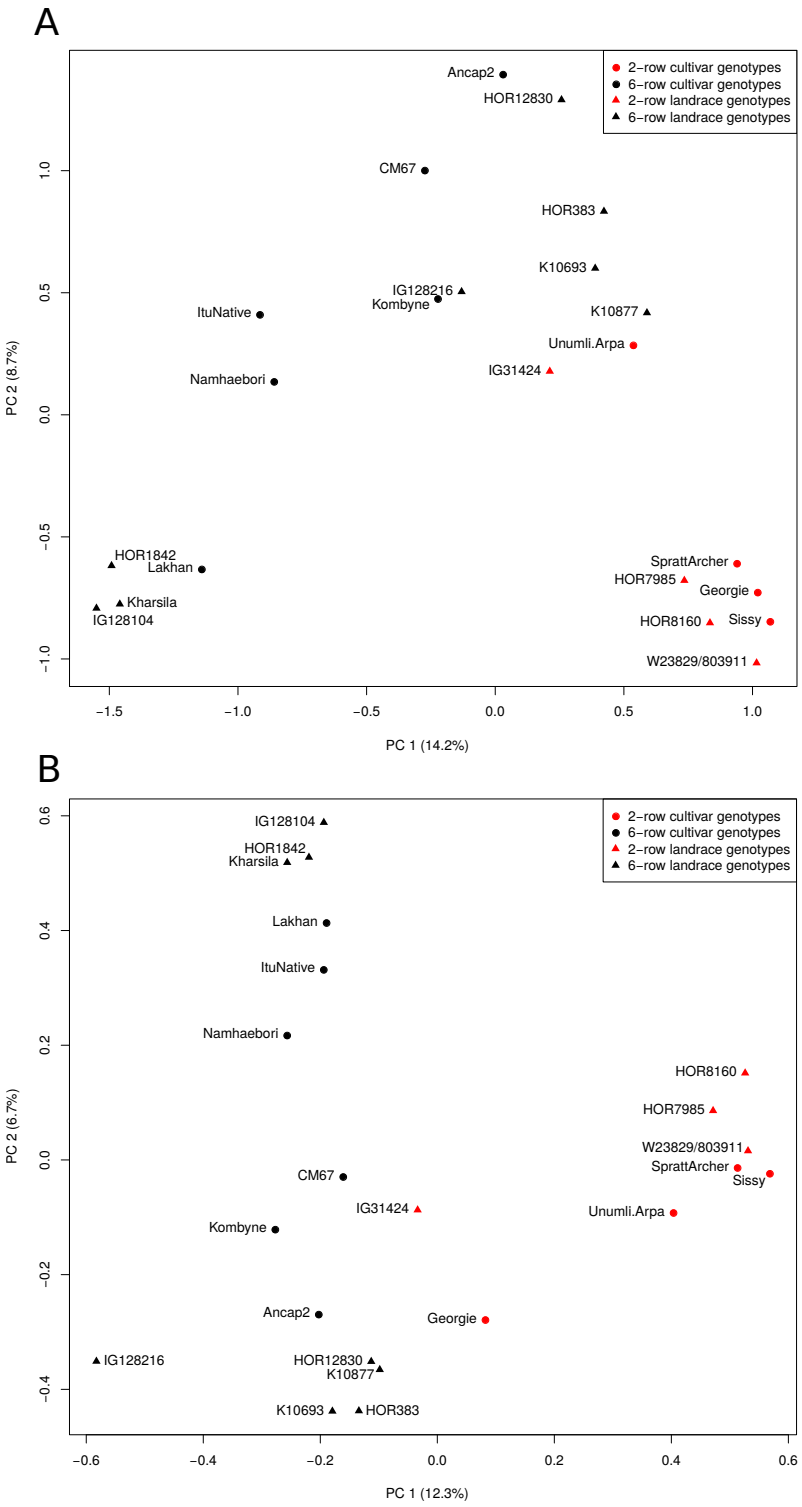

Fig. S3: Principal component analyses of the barley inbred lines considered in our study based on A) 133,566 genome-wide distributed sequence variants, and B) presence/absence allele call at 38,810 expression presence/absence variation. PC 1 and PC 2 are the first and second principal component, respectively, and number in parentheses refer to the proportion of variance explained by the principal components. Symbols identify landrace and cultivar inbreds and colors their row number.

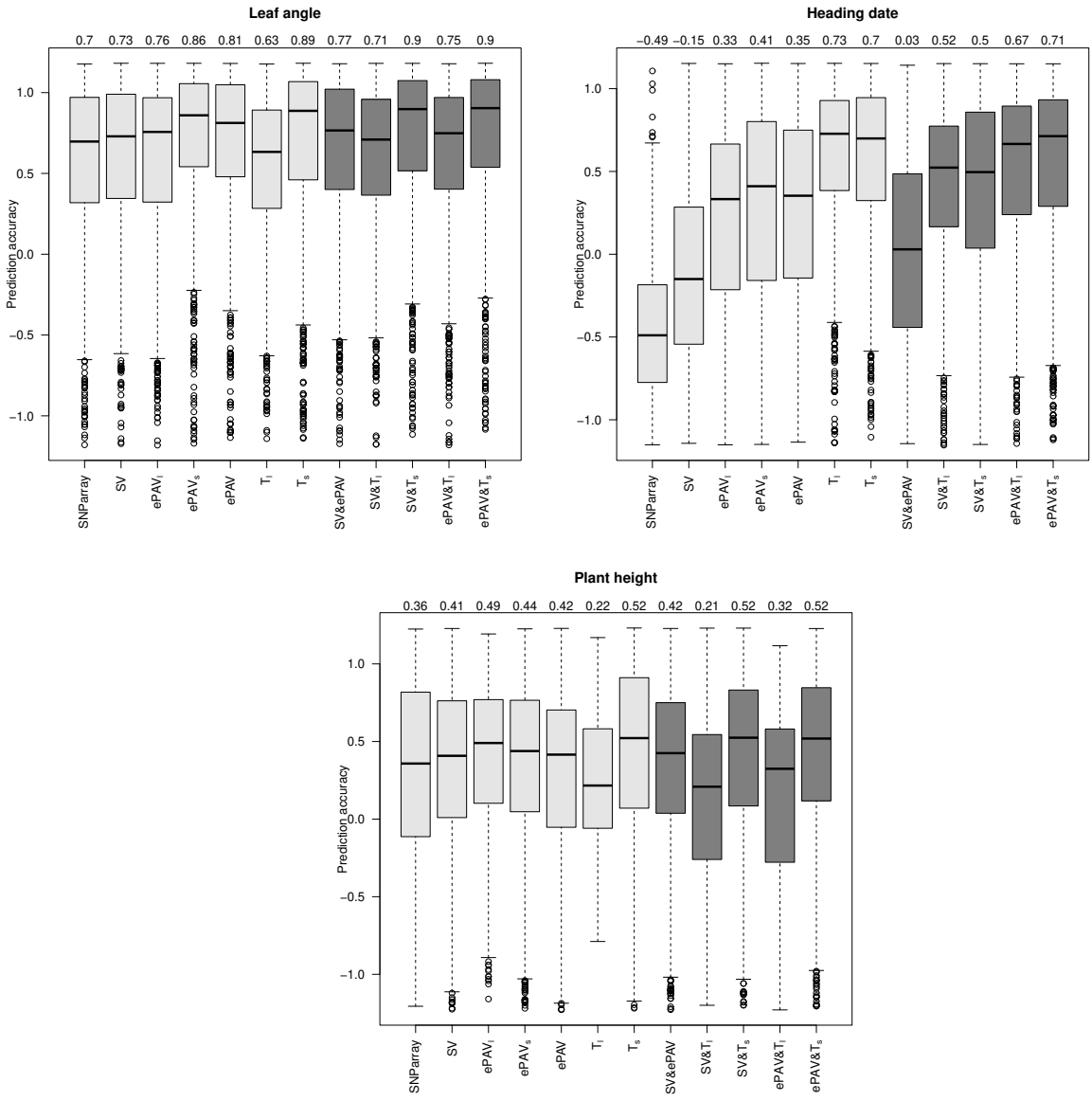

Fig. S4: Prediction accuracy for the barley inbreds of single predictors and combinations thereof for leaf angle, heading date, and plant height from 1,000 cross-validation runs with median prediction accuracy given above each column.

6. Benchmarking of structural variant detection in the tetraploid potato genome using linked-read sequencing

This manuscript is in preparation for submission.

Authors:

Marius Weisweiler, Benjamin Stich

Contribution: First author

Marius Weisweiler and Benjamin Stich designated and coordinated the project.

Marius Weisweiler performed the data analyses.

Marius Weisweiler and Benjamin Stich wrote the manuscript.

Benchmarking of structural variant detection in the tetraploid potato genome using linked-read sequencing

Marius Weisweiler[1] and Benjamin Stich[1,2,*]

[1]Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany.
[2]Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany
[*]Corresponding author: benjamin.stich@hhu.de, Tel: **49-211/81-13395

ABSTRACT

Structural variants (SV) are a potentially important source of phenotypic variation. The objectives of our study were to i) compare the performance of SV callers based on linked-read sequencing to short-read sequencing, ii) examine the influence of SV type, SV length, haplotype incidence (HI), as well as sequencing coverage on the SV calling performance in the tetraploid potato genome, and iii) evaluate the accuracy of detecting insertions by linked-read compared to short-read sequencing. Six linked-read and one short-read SV callers were evaluated based on linked-read sequencing with respect to their precision, sensitivity, and F1-score to detect different SV types with different SV lengths and HIs in the tetraploid potato genome using computer simulations. We observed that Manta and LEVIATHAN reached the maximum precision of SV detection with the highest break point resolution of ≤ 10 bp across all examined SV length categories, whereas LongRanger and VALOR2 showed the lowest. For short SV, high F1-scores averaged across the four HIs were observed for Manta and LinkedSV, whereas for large SV, high F1-scores were observed for the linked-read SV callers. When exploiting linked-read sequencing for SV detection, the vicinity of SV break points provides more signals due to the longer anchor sequences provided by the molecule signals, thereby it is less influenced by the sequencing coverage than using short-read sequencing. Our observations highlighted the importance of short-read signals exploited by Manta and LinkedSV to detect short SV, whereas Manta and NAIBR performed well for detecting larger deletions, inversions, and duplications. Furthermore, insertions can be assembled by Novel-X using linked-read sequencing and, thus, it is superior compared to the detection of insertions based on short-read sequencing.

INTRODUCTION

Structural variants (SV) are commonly defined as genomic rearrangements between individuals or haplotypes that are larger than 49 bp (Ho et al., 2020). SV occur as deletions, insertions, duplications, inversions, or translocations in the genome. In human, SV were tightlier associated with gene expression variation compared to single nucleotide variants (SNV) (Chiang et al., 2017). Additionally, SV were associated with phenotypic variation in several plant species such as wheat and rice (Xu et al., 2012; Li et al., 2012; Nishida et al., 2013). In potato, copy number variation at a limited number of loci was associated with the level of gene expression (Iovene et al., 2013).

Due to the technical improvements of DNA sequencing and novel algorithms (Ho et al., 2020), it is nowadays possible to detect and characterize SV on a genome-wide level. SV detection based on short-read sequencing is well established in human genomics (Cameron et al., 2019; Kosugi et al., 2019) and was also evaluated and used recently for plant genomes (Fuentes et al., 2019; Göktay et al., 2020). However, the reliable detection of SV based on short-read sequencing is challenging due to the necessity of confidently mapped read-pairs (Fang et al., 2019). Additionally, repetitive regions are associated with the occurrence of SV (Hu et al., 2021), where split and paired-end reads can have a low mapping quality due to multi-mapping (Fang et al., 2019). These issues can be avoided by using long-read sequencing (Dierckxsens et al., 2021). However, this approach in turn is associated with high costs and, thus, it is not affordable for many research groups.

Recently, linked-read sequencing was proposed (Weisenfeld et al., 2017; Wang et al., 2019). For linked-read sequencing, paired-end short reads are derived from 50 - 100 kb DNA molecules (Elyanow et al., 2018), which is considerably longer than the read length of most long-read sequencing approaches (cf. Wenger et al., 2019). During the library preparation process, around ten molecules are partitioned into droplets where each DNA fragment (500 bp) derived from these molecules is tagged with a 16 bp long barcode. Due to the random partition of molecules, the likelihood of assigning the same barcode to two molecules from nearby regions in the genome is very low (Elyanow et al., 2018). Therewith, linked-read sequencing provides long-range information as long-read sequencing (Ho et al., 2020) and has the advantages of a high accuracy and low costs as short-read sequencing (Weisenfeld

et al., 2017). However, compared to the established SV detection based on short-read sequencing, less approaches have been described for linked-read based SV calling.

To the best of our knowledge, eight linked-read SV callers were described until today, namely LongRanger (Zheng et al., 2016), GROC-SVs (Spies et al., 2017), NAIBR (Elyanow et al., 2018), ZoomX (Xia et al., 2018), LinkedSV (Fang et al., 2019), Novel-X (Meleshko et al., 2019), VALOR2 (Karaoğlanoğlu et al., 2020), and LEVIATHAN (Morisse et al., 2021a). LongRanger identifies paired-end reads with overlapping barcodes between distant loci. GROC-SVS works similarly to LongRanger with the addition of SV reconstruction using local assemblies. NAIBR exploits discordant paired-end read and split molecule signals in a probabilistic model. ZoomX uses molecule coverage to identify large genomic rearrangements in the human genome. LinkedSV uses short-read signals as read depth, discordance of paired-end reads, and local assembly to detect small deletions. In addition, this tools uses fragments with shared barcodes between two genomic locations and enriched fragment endpoints near break points to detect larger SV (Fang et al., 2019). Novel-X assembles unmapped reads associated with barcodes and maps the resulting contigs to the reference sequence. VALOR2 identifies submolecules using split molecule signals based on barcode information and filters SV candidates using read depth and paired-end read signals. LEVIATHAN identifies a number of shared barcodes in specific regions and secondly, discordant paired-end and split read signals are then used to filter SV candidates (for review see Ho et al., 2020).

With the exception of LEVIATHAN, all of the above mentioned SV callers were up to now only evaluated for SV detection in the human genome. LEVIATHAN was also evaluated for SV detection in the butterfly (*H. numata*) genome (Morisse et al., 2021a). To our knowledge, no study is available where SV detection using linked-read sequencing is evaluated for plant species despite the differences between the plant and human genome with respect to genome size, repeat content, or ploidy. Furthermore, to our knowledge, it is also the first study where SV calling is evaluated for a polyploid genome.

Therefore, the objectives of our study were to i) compare the performance of SV callers based on linked-read sequencing to short-read sequencing, ii) examine the influence of SV type, SV length, haplotype incidence (HI), as well as sequencing coverage on the SV calling performance in the tetraploid potato genome, and iii) evaluate the accuracy of detecting

insertions by linked-read compared to short-read sequencing.

MATERIAL AND METHODS

Simulation preparation and genome mutation

We used Mutation-Simulator (version 2.0.3) (Kühl et al., 2021) to simulate deletions, duplications, inversions, and insertions in the first and second chromosome of the dAg1_v1.0 potato reference sequence (Freire et al., 2021) which is a consensus sequence of the two haplotypes of a diploid clone derived from the commercially important potato variety Agria. We considered five SV length categories for each of the above mentioned SV types (A: 50 - 300 bp; B: 0.3 - 5 kb; C: 5 - 50 kb; D: 50 - 250 kb; E: 0.25 - 1 Mb). Mutation-Simulator was used with the mutation rates of $7.0\text{x}10^{-6}$ ($\sim$ 800 - 1000 SV) for the SV length categories A - C, $7.0\text{x}10^{-7}$ ($\sim$ 90 SV) for D, and $3.5\text{x}10^{-7}$ ($\sim$ 45 SV) for E.

In a first step, simulations on a homozygous level were performed where the SV were present in all four haplotypes (4/4) of the simulated potato genome. In addition to the homozygous level, we simulated heterozygous SV with HIs of one to three (if SV occurs in one, two, or three haplotypes). To do this, a custom python script was used to prepare heterozygous SV for simulations, where the SV was only present in one of the four haplotypes (1/4). Which of the four haplotypes received the SV was randomly determined for each SV. The same procedure was used to simulate SV in two out of four (2/4) as well as three out of four (3/4) haplotypes. For each heterozygous SV simulation, the total number of simulated SV corresponded to that of the above described homozygous simulation of the specific SV type and SV length category combination. The identification of the correct HI by the SV callers was not possible because polyploid genotyping algorithms were not implemented in these SV callers. Simulations for each SV type* SV length category* HI combination were replicated five times.

In addition to the simple simulations explained above, where the SV types, SV length categories, and HIs were simulated separately, we performed complex simulations (Fig. 1). In these complex simulations, different SV types, SV length categories, and HIs were simulated together to mimic more closely experimental potato genome sequences. Additionally, 80,000 single nucleotide variants (SNV) and 600 small insertions and deletions (INDELs, 2

- 49 bp) were included. The numbers of SV for each SV type (464 deletions, 464 insertions, 124 duplications, 108 inversions) and SV length category were chosen based on the average number of SV observed in experimental data for 100 tetraploid potato clones (Weisweiler and Baig et al., in preparation). For each SV type and SV length category, 25% of SV were simulated for each of the four different HIs. The complex simulations were replicated 20 times.

Linked-read simulation and mapping

LRSim (version 1.0) (Luo et al., 2017) was used to simulate linked reads (-f 50 -t 20 -m 10) with a sequencing coverage of 45x, 90x, 135x, and 180x resulting in a sequencing coverage per haplotype of about 11x, 22x, 34x, and 45x, respectively. The mean molecule size was set to 50 kb, the molecules per partition to 10 and the number of partitions to 20,000 as it was recommended by Luo et al. (2017) for *Arabidopsis thaliana* which have a similar genome size as the first two chromosomes of the dAg1_v1.0 reference sequence (Freire et al., 2021). Linked reads were mapped against the non-mutated dAg1_v1.0 reference sequence with LongRanger wgs (version 2.2.2).

SV calling and filtering

LRez (version 2.2.2) (Morisse et al., 2021b) was used to index bam files for LEVIATHAN. Sonic (version 1.2) (`https://github.com/calkan/sonic/`) was used to create the sonic file for VALOR2. The simulated SV were called using Manta (version 1.6) (Chen et al., 2016) as benchmark short-read SV caller. In addition, LEVIATHAN (-v 50, version 1.0.1), LinkedSV (--wgs --germline_mode, gap regions, version 1.0.1), VALOR2 (sonic file, -p 4, -c 2, version 2.1.5), LongRanger wgs (version 2.2.2), Novel-X (version 0.3) (Meleshko et al., 2019), and NAIBR (Elyanow et al., 2018) were evaluated as linked-read SV callers (Table 1). Additionally, LinkedSV and LongRanger can detect small deletions based on short-read sequencing signals. This was indicated in the following as LinkedSV (short) and LongRanger (short). All SV callers, independent from the usage of short-read or linked-

read signals, were evaluated based on simulated linked-read sequencing data. The workflow described above was implemented in Snakemake (version 5.10.0) (Köster et al., 2021) and is available via github (`https://github.com/mw-qggp/SV_simulation_potato`).

In the next step, the detected SV were filtered. A SV call was only kept if it passed the built-in filters of the respective SV caller. SV calls which were annotated as "BND" were filtered out. SV calls which covered regions in the reference sequence consisting of N's were filtered out as well. Additionally, for some SV callers additional filter criteria were applied: for LongRanger, SV calls with the annotation "UNK", which is defined as unknown SV type, were not considered. Additionally, for LinkedSV and Manta where each inversion was called twice, only one inversion entry was kept to avoid incorrect statistics. For NAIBR, the orientation of novel adjacencies was used as SV type annotation.

Evaluation of SV calling

We calculated the sensitivity (1), precision (2), and the F1-score (harmonic average of the precision and sensitivity) (3) as

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

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

$$F1 - score = 2 * (Precision * Sensitivity/Precision + Sensitivity) \tag{3}$$

for all combinations of SV types* SV callers* HIs, where TP was the number of true positive SV, FP the number of false positive SV, and FN the number of false negative SV. Before calculating the above described evaluation criteria, the break point resolution (BPR) for each SV length category was estimated for all SV callers based on 135x sequencing coverage for all SV types. Based on this analysis, the following BPR thresholds were chosen to allow a fair comparison between the SV callers (Supplementary Table S1). For SV length category A, a TP SV had break points that did not differ more than 10 bp from those of the simulated SV and the SV length did not differ by more than 10 bp. For the SV length category B, a TP SV had break points and length differences compared to the simulated SV of $\leq$ 50 bp. For the SV length category C, a TP SV had break points and

length differences compared to the simulated SV of $\leq$ 160 bp. For duplications of the SV length categories D and E, a TP SV had break points and length differences compared to the simulated SV of $\leq$ 250 bp. For deletions and inversions of the SV length category D, $\leq$ 550 bp and $\leq$ 800 bp were chosen as threshold, respectively. For deletions and inversions of the SV length category E, $\leq$ 250 and $\leq$ 550 bp were used, respectively. For insertions, the start of a TP insertion had a break point that did differ $\leq$ 10 bp from the start of the simulated insertion to allow a fair comparison between Manta and Novel-X due to the absence of an insertion length for Manta. Additionally, for Novel-X, called insertions were also evaluated considering two break points as it was done for deletions to determine the precision of the detected insertion length. The sequence similarity between detected and simulated insertions was evaluated. This was realized by pairwise alignments using stretcher from the EMBOSS package (version 6.6.0.0) (Rice et al., 2000).

For each TP SV, the called SV had to be annotated as the considered SV type. For deletions and duplications called by LEVIATHAN, the SV type annotation was ignored in a second evaluation (LEVIATHAN (IG)), because pre-simulations have shown that a bug in the algorithm of LEVIATHAN makes it difficult to differ between deletions and duplications. To determine the final sensitivity and precision values, as well as the final F1-scores for the simple and complex simulation scenarios, the median across the five (simple) as well as 20 (complex) replications was calculated. We only evaluated the performance of SV callers for the SV length categories C - E for the complex simulations. For the detection of insertions in the complex simulations, all SV length categories were evaluated together because detected insertions could not be separated by the SV length category for Manta.

RESULTS

Six linked-read and one short-read SV callers (Table 1) were evaluated based on linked-read sequencing with respect to their precision, sensitivity, and F1-score to detect different SV types with different SV lengths and HIs in the tetraploid potato genome using computer simulations.

BPR of SV callers

In a first step, the BPR of each SV caller was determined for the detection of homozygous (4/4) deletions (insertions for Novel-X) for each SV length category based on a 135x sequencing coverage. Deletions have been chosen as SV type and 135x as sequencing coverage, because all SV callers, except VALOR2 and LEVIATHAN, have been developed to detect deletions of all SV length categories.

We observed considerable differences among the BPR of the different SV callers (Fig. 2). Across all examined SV length categories, Manta and LEVIATHAN reached the maximum precision of SV detection with the highest BPR of $\leq$ 10 bp. In contrast, the BPR of LongRanger and VALOR2 were the lowest.

The trends of the BPR observed for the other SV types corresponded well to those observed for deletions (Supplementary Fig. S1, S2). The main exception was VALOR2, where BPR were observed for large inversions that were even lower than the BPR of deletions.

SV detection for different SV length categories

First, we focused on the detection of SV based on a sequencing coverage of 135x which corresponds to that of an experimental study with about 100 tetraploid potato clones (Weisweiler and Baig et al. in preparation).

All SV callers, except Novel-X, were able to detect deletions for at least one SV length category. For the SV length categories A and B, the highest F1-scores averaged across the four HIs (hereafter designated as average F1-score) were observed for Manta with 98.3% (for A

and B) followed closely by LinkedSV (short) (95.9%, 95.6%, Fig. 3III), and with a considerable difference by LongRanger (short) (23.4%, 22.5%). Linked-read SV callers without an implemented short-read algorithm were not able to detect deletions of the SV length category A and B (Supplementary Table S2, S3). Larger deletions could be identified by linked-read SV callers (Supplementary Table S4, S5, S6). However, for the SV length category C, the average F1-scores of Manta with 98.2% and LinkedSV (short) with 92.6% were still higher compared to those of the SV callers without an implemented short-read algorithm. The highest F1-score of a linked-read SV caller was observed for LEVIATHAN (IG) with an average F1-score of 88.0%. For the SV length categories D, increased average F1-scores were observed for the linked-read SV callers as for NAIBR (92.9%) and Longranger (linked) (87.3%), whereas a decreased average F1-score was observed for LinkedSV (short) (43.1%). For the SV length category E, a similar figure was observed, where Manta (89.6%) and NAIBR (88.5%) showed the highest average F1-scores.

The performance of detecting inversions showed a similar trend as it was observed for deletions. For the SV length categories A and B, the short-read SV caller Manta performed well with high average F1-scores (90.0%, 98.9%) (Fig. 4III, Supplementary Table S7, S8), whereas linked-read SV callers, especially LEVIATHAN (91.4%), showed high average F1-scores for larger inversions of the SV length category C. Additionally, the average precision values were very high for LinkedSV (99.4%) and NAIBR (98.3%) (Supplementary Table S9). An even better performance of linked-read SV callers was observed for the SV length categories D and E (Supplementary Table S10, S11), especially for NAIBR and LEVIATHAN.

With the exception of VALOR2, the same SV callers which could detect inversions were able to detect duplications. As it was observed for deletions and inversions, Manta was the best SV caller to identify duplications for the SV length categories A with an average F1-score of 66.2% (Fig. 5III) which was considerably lower compared to those values for calling deletions (98.3%) and inversions (90.0%). This is caused by a low sensitivity (58.6%) rather than by a low precision (82.2%) (Supplementary Table S12). LEVIATHAN (IG) was the only linked-read SV caller which could detect duplications of the SV length category B, but the average F1-score, sensitivity, and precision values were with 6.4%, 3.5%, and 52.6%, respectively, considerably lower compared to those values observed for

Manta (97.7%, 95.7%, 99.8%) (Supplementary Table S13). For the SV length category C, Manta performed well with an average F1-score of 97.2%, followed by LEVIATHAN (IG) (84.4%). LongRanger showed a considerably lower F1-score of 34.4% because of the low sensitivity (21.8%) (Supplementary Table S14). In contrast to the SV length category C, NAIBR and LinkedSV were able to detect duplications of the SV length category D (Supplementary Table S15). Manta, NAIBR, and LongRanger performed well with average F1-scores ranging from 88.9 to 92.6%. For the SV length category E (Supplementary Table S16), the highest average F1-scores were observed for Manta (85.2%) and NAIBR (85.3%). Manta and Novel-X were the only two SV callers that were able to detect insertions. Manta as short-read SV caller could detect the break point of the insertion start position but could not assemble the inserted sequence. Therefore, the performance of Manta and Novel-X was compared based on one break point at the insertion start position. For the SV length category A, Manta showed considerably higher F1-scores (94.5 - 99.5%) for all four HIs compared to Novel-X (45.7 - 87.6%) (Fig. 6III). The precision of Novel-X to detect insertions of the SV length category A was with values between 98.2 and 98.9% high, but the sensitivity was low (29.6 - 78.7%) (Supplementary Table S17). For the SV length categories B and C, Novel-X performed with F1-scores between 97.3 and 98.6% better than Manta (86.7 - 99.2%) for almost all four HIs. In addition to the comparison of Manta and Novel-X, the performance of Novel-X was also evaluated as it was done before for the other SV types to determine the precision to assemble the inserted sequence. With exception of the SV length category E, the evaluation of Novel-X based on two break points has shown similar F1-scores compared to the evaluation based on only one break point (Supplementary Tables S18 - S21).

SV detection based on different sequencing coverages

Apart from the influence of the SV type and SV length on the SV calling performance, we examined the influence of the sequencing coverage. To do so, four different sequencing coverages, namely 45x, 90x, 135x, and 180x were considered.

The performance to detect deletions of the short-read SV callers increased with increasing sequencing coverage (Fig. 3, Supplementary Tables S2 - S6). This was especially true for

the detection of deletions of the SV length category A and B. The F1-score of Manta e.g. increased from 81.1% (45x) to 98.1% (180x) for the detection of deletions of the SV length category A and the HI 1/4. Even higher was the difference for this simulation for LinkedSV (short) with an increase of 50.3%. This strong influence of the sequencing coverage on the F1-score was not observed for the detection of inversions and duplications of the SV length categories A and B.

Linked-read SV callers, especially NAIBR and LinkedSV (linked) performed more independently from the sequencing coverage than short-read SV callers (Fig. 3 - 5). The only exception was the detection of insertions. The average F1-scores of Novel-X increased considerably with an increasing coverage (Fig. 6).

SV detection using different HIs

We also examined the role of HIs on the performance of SV detection. In most of the simulation scenarios, a higher F1-score was observed for the simulations of the HI 1/4 and 4/4 compared to 2/4 and 3/4 scenarios. This was especially true for the SV length categories D and E for all SV types and for the SV callers Manta and NAIBR. Exceptions of this trend were the performance of LinkedSV (linked) and LEVIATHAN (IG) for the detection of deletions and duplications of the HI 1/4 and NAIBR for the detection of deletions and inversions of the SV length category C. Further, Novel-X showed a higher F1-score to detect insertions of the SV length category A for the HI 2/4 and 4/4 compared to 1/4 and 3/4. Interestingly, the performance of VALOR2 was more independent from the HI compared to the other SV callers.

Evaluation of SV detection using complex simulations

In addition to the simple simulations, where the combinations of SV types, SV length categories, as well as HIs were simulated separately, we performed complex simulations including all features of the simple simulations together to mimic experimental potato genome sequencing data.

In general, the F1-scores observed in the complex simulations showed a high accordance to the results of the simple simulations. For the detection of the different SV types, Manta and NAIBR showed sensitivity and precision values up to 100.0% for most of the SV length categories for all sequencing coverages (Tables 2 - 5). In contrast to the simple simulations, LongRanger (linked) showed lower sensitivity values for the detection of larger deletions.

DISCUSSION

Due to tremendous improvements of sequencing technologies and bioinformatic tools, genome-wide SV detection became possible in the last years (Ho et al., 2020). Algorithms based on short-read and long-read sequencing were developed to detect SV. However, despite well established SV detection based on short-read sequencing in the human genome (Cameron et al., 2019; Kosugi et al., 2019), low precision and a lack of detecting large SV as well as assembling insertions were reported (Chaisson et al., 2015; Huddleston and Eichler, 2016; Meleshko et al., 2019; Ho et al., 2020). In contrast, SV calling based on long-read sequencing overcomes these issues but higher operational costs, large DNA input requirement, as well as lower sample throughput (Ho et al., 2020) are the consequences. We therefore benchmarked in a plant genome context SV callers which were developed to detect SV based on linked-read sequencing, as the latter has the potential to exploit signals of short-read sequencing and long-range information. Two previously described linked-read SV callers were not considered in our study, due to discontinued support (GROC-SVs) (Spies et al., 2017) or the restriction to human genomes (ZoomX) (Xia et al., 2018).

Simple vs. complex simulations

In general, the high sensitivity and precision values observed in the simple simulations (Figures 3 - 6, Supplementary Tables S2 - S21) could be confirmed in the complex simulations (Tables 2 - 5). Therefore, only the results of the simple simulations were discussed in the following. In both simulation scenarios, maximum precision values of 100% were frequently observed for all SV types and SV length categories. This finding suggests that the different SV types and SV lengths have no negative influence on the detection of each other and, thus, the high precision values observed in our complex simulations can be also expected in experimental data of tetraploid potato varieties.

SV detection based on short-read vs. linked-read signals

The linked-read sequencing data simulated in our study can be used to evaluate SV detection based on short-read and linked-read signals. The linked-read signals are, except for the mapping of the reads, simply not considered by the short-read SV callers to call SV.

We observed high precision and sensitivity values for the SV detection using the short-read algorithms implemented in Manta and LinkedSV (short) (Fig. 3, Supplementary Tables S2 - S6). Our observations are supported by recent comprehensive SV calling evaluation studies in humans (Cameron et al., 2019; Kosugi et al., 2019). However, our figures are in contrast to the low precision of around 15% and sensitivity values between 30 and 70% which have been frequently reported for the detection of SV based on short-read sequencing in the context of the human genome (English et al., 2015; Sudmant et al., 2015; Sedlazeck et al., 2018; Sethi et al., 2020). One reason might be that the latter studies evaluated SV callers that have been developed ten years ago such as Pindel (Chen et al., 2009) or BreakDancer (Abyzov et al., 2011). These SV callers only exploit one single short-read signal whereas the nowadays available tools use a combination of read depth, paired-end reads, and split reads to increase the sensitivity and precision (Weisweiler et al. 2022 in review). An additional reason for the high precision and sensitivity observed in our study might be the improved accuracy of read mapping by considering the linked-read information for that step of the analysis (Marks et al., 2019).

In our study, the F1-score of the short-read SV caller Manta was always equal or higher compared to that of the linked-read SV callers NAIBR or LinkedSV (Fig. 3 - 6), whereas in Fang et al. (2019), these linked-read SV callers showed higher F1-scores than those of the short-read SV callers Lumpy (Layer et al., 2014) and Delly (Rausch et al., 2012). This observation can be explained thereby that Manta showed a better performance to detect SV in human (Cameron et al., 2019; Kosugi et al., 2019) and barley (Weisweiler et al. 2022 in review) compared to Delly and Lumpy. The lower F1-score of linked-read SV callers is caused by a lower sensitivity of the linked-read SV callers compared to Manta. In contrast, the precision was high for short- and linked-read SV callers (Supplementary Tables S2 - S21). The high precision of linked-read SV callers can be explained by the usage of short-read signals and barcode information which was also previously reported in human data

sets (Sethi et al., 2020). Due to the usage of additional information provided by linked-read sequencing, linked-read SV callers should be able to increase the sensitivity. However, the lower sensitivity of linked-read SV callers compared to Manta indicates that linked-read SV callers cannot use all information provided by linked-read sequencing. A reason for this might be the relatively recent history and the corresponding low level of elaboration of linked-read compared to short-read SV calling algorithms (Sethi et al., 2020).

Our finding indicates that further improvements are possible for linked-read SV callers. Furthermore, the combination of short-read signals and long-range information based on molecule signals is expected to increase the precision of SV detection. Therefore, until improved linked-read SV callers are available, we suggest the combined usage of both, short-read and linked-read SV callers, based on linked-read sequencing data to maximize the sensitivity but retaining a high precision.

Influence of SV length on SV detection and performance of SV callers

In order to being able to interpret properly the observed numbers of detected SV of different SV lengths and SV types in experimental studies, detailed knowledge about the sensitivity and precision of SV callers for different SV length categories is required.

Except for insertions, linked-read SV callers were not able to detect SV of the SV length category A (50 - 300 bp) and B (0.3 - 5 kb) or the performance was on a low level (e.g. LEVIATHAN) (Fig. 3, 4, 5). In contrast, Manta as short-read SV caller as well as the short-read algorithm of LinkedSV performed well for these SV length categories. The linked-read SV callers were developed for the detection of large SV ($\geq$ 10 kb) (Zheng et al., 2016; Fang et al., 2019) and the focus did not lay on the detection of small SV. However, NAIBR and LEVIATHAN were able to detect SV between 1 - 5 kb in the human genome, even though showing a low sensitivity (Elyanow et al., 2018; Morisse et al., 2021a) which is in agreement wih our results for LEVIATHAN. The reason for the discrepancy of SV detection by NAIBR remains elusive. An obvious reason for the low performance of linked-read SV callers to detect short SV in our study is that the principle of SV detection based on linked-read barcode information is not suitable here. The specific signals of linked-read SV calling as overlapping barcodes or split molecules cannot be used because of the short

distance between the two break points of a short SV. Therefore, these SV can only be detected based on short-read signals as discordant paired-end reads, split reads, or unusual read depth.

The sensitivity and precision of the linked-read SV callers to detect SV of the SV length categories C - E (5 kb - 1 Mb) for all SV types was considerably higher compared to the SV length category A and B (Supplementary Tables S2 - S21). In addition, Manta performed also well for large SV for all SV types. Our results were supported by a previous study in human, where a high precision of NAIBR and LinkedSV and a considerably lower precision of LongRanger was reported for the detection of large SV (Fang et al., 2019). The high precision to detect large deletions and inversions in the human genome reported for VALOR2 (Karaoğlanoğlu et al., 2020) could be supported by our results as well (Supplementary Tables S5, S6, S10, S11). However, these come together with the costs of a lower sensitivity and a considerably lower BPR compared to that of the other SV callers (Fig. 2, Supplementary Fig. S1, S2).

Influence of sequencing coverage on SV detection

First, we assessed the influence of the sequencing coverage on the performance of short-read algorithms based on linked-read sequencing. The strongest differences were observed for calling deletions of the SV length category A (Fig. 3, Supplementary Table S2) when increasing the sequencing coverage from 45x (∼11x per haplotype in potato) to 90x (∼22x per haplotype), where the sensitivity increased by 23.3% for Manta and 45.6% for LinkedSV (short). This trend was also observed for the other SV length categories albeit in alleviated terms. Further, the performance of short-read algorithms increased only marginally when increasing the sequencing coverage to 135x and 180x, respectively. Our observations are in accordance with results of Cameron et al. (2019) who reported a higher sensitivity for short-read SV callers using higher levels of sequencing coverage. These findings can be explained by the fact that short-read sequencing with higher coverage results in an increased number of short-read signals such as discordant paired-end and split reads (Kosugi et al., 2019). This in turn results in a higher sensitivity.

In contrast to the SV detection based on short-read signals, the influence of sequencing

coverage on the performance of linked-read SV callers was marginal (Fig. 3 - 5). The good performance of linked-read SV callers independent from the sequencing coverage can be explained by additional signals comprised in linked-read sequencing data sets which are created during the library preparation process. When exploiting linked-read sequencing for SV detection, the vicinity of SV break points provides more signals due to the longer anchor sequences provided by the molecule signals. In contrast, for short-read sequencing, only reads can be considered for the SV detection where the sequence covered the break points. Therefore, the reduction of the sequencing coverage results in less short-read signals which has more severe consequences for the SV detection compared to linked-read signals. In contrast to the above described trend for linked-read SV callers, we have observed two exceptions where the sequencing coverage influenced the SV detection for linked-read SV callers. First, detecting insertions by Novel-X was strongly influenced by the sequencing coverage (Fig. 6). An insufficient coverage leads to difficulties in reassembling the anchor sequences for the detected insertions and, thus, the break points of the insertions cannot be determined (Meleshko et al., 2019). Second, SV detection for the SV length category C of the HI 1/4 scenario by LEVIATHAN (IG) was strongly influenced by the sequencing coverage e.g. for deletions (40.1%) (Supplementary Table S4) or inversions (20.4%) (Supplementary Table S9). An explanation for the weak performance of LEVIATHAN (IG) for calling SV for the HI 1/4 scenario on 45x sequencing coverage could be that after considering the barcode information, short-read signals such as discordant paired-end or split reads are used to process candidate SV (Morisse et al., 2021a). However, as explained above, short-read signals are strongly influenced by sequencing.

Influence of HI on SV detection in a tetraploid genome

We examined the performance of SV callers using different HIs for the tetraploid potato genome. As expected, the performance of all SV callers was better for simulation scenarios with a HI 4/4 than for the other HI scenarios. However, the observed performance for the HIs 2/4 and 3/4 was worse compared to those for the HI 1/4 (Fig. 3 - 6). The reason for this observation remains elusive and additional research is needed in the field of polyploid SV calling.

We observed that the different HI scenarios influenced the performance of the SV callers, but we were not able to evaluate the correct identification of the HI by the SV callers due to a lack of implemented genotyping algorithms. Approaches for SV genotyping based on short-read sequencing have been described for diploid genomes (Hickey et al., 2020) even though it is more complex (Cameron et al., 2017) compared to well established SNV genotyping based on read depth signals (Poplin et al., 2017). Recently, it has been shown that SNV genotyping is more error-prone for polyploid than for diploid genomes with the request of attention interpreting polyploid genotype calls and a need for further improvements (Cooke et al., 2022). Considering the need of improvements of diploid SV genotyping (Chander et al., 2019; Khayat et al., 2021) and the issues of polyploid SNV genotyping (Cooke et al., 2022), polyploid SV genotyping will be one of the big challenges in crop research.

Assembling insertions using linked-read sequencing

An obvious drawback of SV calling using short-read sequencing is the lack of detecting larger insertions ($\geq$ 0.3 kb) (Rizk et al., 2014; Holtgrewe et al., 2015; Kehr et al., 2016; Kavak et al., 2017) caused by the limited anchor size due to the small insert size of the sequencing library and the corresponding incapacity to span over larger repetitive regions in the genome (Meleshko et al., 2019). Manta is able to determine the length for insertions up to $\sim$1 kb. SV calling using linked-read sequencing can principally raise this threshold. However, up to date, only one algorithm (Novel-X) was developed for the detection of insertions using linked-read sequencing data.

As this algorithm revealed high sensitivity and precision values to detect insertions (Fig. 6, Supplementary Tables S17 - S21), we evaluated the assembled length of the insertions. Considering both break points to determine the length of the insertions, high sensitivity and precision values were observed for Novel-X. Furthermore, we observed sequencing similarities of 100% between five simulated and detected insertions for each SV length category. This observation was in accordance to Meleshko et al. (2019) who reported similar values for the human genome. These observations illustrate the potential of linked-reads and especially of Novel-X to detect and assemble insertions.

Computational performance of SV callers

To compare the computational performance of the different SV callers, we examined the resources needed by SV callers in the case of 180x sequencing coverage in the complex simulations (Table 6). We have observed a short CPU time and low memory requirement for Manta compared to the considerably higher values for the linked-read SV callers. High memory peaks as observed for LEVIATHAN could lead to issues when SV calling is examined on a whole genome level for species with large genomes.

Conclusion

We observed high precision and sensitivity values for SV detection in the potato genome. Our observations highlighted the importance of short-read signals exploited by Manta and LinkedSV to detect short SV, whereas Manta and NAIBR performed well for detecting larger deletions, inversions, and duplications. Furthermore, we illustrated that large insertions can be assembled by Novel-X using linked-read sequencing and, thus, it is superior compared to the detection of insertions based on short-read sequencing. The BPR was similar for the different SV types, where we observed the highest BPR for Manta and LEVIATHAN. The HI influenced the performance of all SV callers, where for the HI 4/4 and 1/4 scenarios, the highest precision and sensitivity values were observed. Finally, the short-read algorithms were more strongly influenced by the sequencing coverage than the linked-read SV callers, except Novel-X, where at least a sequencing coverage of 90x should be used to detect insertions.

DECLARATIONS

Availability of data and materials

Snakemake workflows of the simple and complex simulations are available via github (`https://github.com/mw-qggp/SV_simulation_potato`). Further scripts are available from the authors upon request.

Acknowledgements

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

Funding

This research was funded by the Federal Ministry of Food and Agriculture/Fachagentur Nachwachsende Rohstoffe (grantID 22011818, PotatoTools). 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.

Authors' contributions

MW and BS designed the project; MW performed the analyses; MW and BS wrote the manuscript. All authors read and approved the final manuscript and they declare that they have no competing interests.

REFERENCES

Abyzov A, Urban AE, Snyder M, Gerstein M (2011), CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21:974–984

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

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 Research 27:1–11

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, et al. (2015), Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611

Chander V, Gibbs RA, Sedlazeck FJ (2019), Evaluation of computational genotyping of structural variation for clinical diagnoses. GigaScience 8:1–7

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, et al. (2009), BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6:677–681

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, et al. (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, et al. (2017), The impact of structural variation on human gene expression. Nature Genetics 49:692–699

Cooke DP, Wedge DC, Lunter G (2022), Benchmarking small-variant genotyping in polyploids. Genome Research 32:403–408

Dierckxsens N, Li T, Vermeesch J, Xie Z (2021), A benchmark of structural variation detection by long reads through a realistic simulated model. Genome Biology 22:342

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

English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, et al. (2015), Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Genomics 16:286

Fang L, Kao C, Gonzalez MV, Mafra FA, Pellegrino da Silva R, Li M, Wenzel SS, et al. (2019), LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. Nature Communications 10:5585

Freire R, Weisweiler M, Guerreiro R, Baig N, Hüttel B, Obeng-Hinneh E, Renner J, et al. (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, et al. (2019), Structural variants in 3000 rice genomes. Genome Research 29:870–880

Göktay M, Fulgione A, Hancock AM (2020), A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. Molecular Biology and Evolution 38:1498–1511

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, et al. (2020), Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biology 21:35

Ho SS, Urban AE, Mills RE (2020), Structural variation in the sequencing era. Nature Reviews Genetics 21:171–189

Holtgrewe M, Kuchenbecker L, Reinert K (2015), Methods for the detection and assembly of novel sequence in high-throughput sequencing data. Bioinformatics 31:1904–1912

Hu Y, Colantonio V, Müller BS, Leach KA, Nanni A, Finegan C, Wang B, et al. (2021), Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nature Communications 12:1227

Huddleston J, Eichler EE (2016), An incomplete understanding of human genetic variation. Genetics 202:1251–1254

Iovene M, Zhang T, Lou Q, Buell CR, Jiang J (2013), Copy number variation in potato - an asexually propagated autotetraploid species. Plant Journal 75:80–89

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

Kavak P, Lin YY, Numanagić I, Asghari H, Güngör T, Alkan C, Hach F (2017), Discovery and genotyping of novel sequence insertions in many sequenced individuals. Bioinformatics 33:i161–i169

Kehr B, Melsted P, Halldórsson BV (2016), PopIns: population-scale detection of novel sequence insertions. Bioinformatics 32:961–967

Khayat MM, Mohammad S, Sahraeian E, Zarate S, Carroll A, Hong H, Pan B, et al. (2021), Hidden biases in germline structural variant detection. Genome Biology 22:347

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. (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 Biology 20:117

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 Biology 15:R84

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

Luo R, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC (2017), LRSim: a linked-reads simulator generating insights for better genome partitioning. Computational and Structural Biotechnology Journal 15:478–484

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, et al. (2019), Resolving the full spectrum of human genome variation using linked-reads. Genome Research 29:635–645

Meleshko D, Marks P, Williams S, Hajirasouliha I (2019), Detection and assembly of novel sequence insertions using Linked-Read technology. bioRxiv https://doi.org/10.1101/551028

Morisse P, Legeai F, Lemaitre C (2021a), 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

Morisse P, Lemaitre C, Legeai F (2021b), LRez: C ++ API and toolkit for analyzing and managing Linked-Reads data. arXiv 2103.14419v2

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. Molecular Breeding 31:27–37

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, et al. (2017), Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 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

Rice P, Longden I, Bleasby A (2000), EMBOSS: the european molecular biology open software suite. Trends in Genetics 16:276–277

Rizk G, Gouin A, Chikhi R, Lemaitre C (2014), MindTheGap: integrated detection and assembly of short and long insertions. Bioinformatics 30:3451–3457

Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018), Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics 19:329–346

Sethi R, Becker J, de Graaf J, Löwer M, Suchan M, Sahin U, Weber D (2020), Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions. PLOS Computational Biology 16:e1008397

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, et al. (2017), Genome-wide reconstruction of complex structural variants using read clouds. Nature Methods 14:915–920

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, et al. (2015), An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Wang O, Chin R, Cheng X, Yan Wu MK, Mao Q, Tang J, Sun Y, et al. (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 Research 29:798–808

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

Weisweiler M, Arlt C, Wu PY, Van Inghelandt D, Hartwig T, 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. PLOS Genetics In review

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, et al. (2019), Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37:1155–1162

Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP (2018), Identification of large rearrangements in cancer genomes with barcode linked reads. Nucleic Acids Research 46:e19

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, et al. (2012), Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology 301:105–111

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

Table 1: Properties of structural variant (SV) callers.

		Detection of			
SV caller	Detection mode	Deletions	Insertions	Inversions	Duplications
Manta	short	x	x	x	x
LinkedSV	short + linked	x		x ($\geq$ 10kb)	x ($\geq$ 20kb)
LongRanger	short + linked	x		x ($\geq$ 30kb)	x ($\geq$ 30kb)
VALOR2	linked	x ($\geq$100kb)		x ($\geq$ 80kb)	
NAIBR	linked	x		x	x
LEVIATHAN	linked	x ($\geq$ 1kb)		x ($\geq$ 1kb)	x ($\geq$ 1kb)
Novel-X	linked		x		

Table 2: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length categories C (5 - 50 kb), D (50 - 250 kb), and E (0.25 - 1 Mb) in complex simulations.

	Sequencing coverage			
SV caller	45x	90x	135x	180x
	SV length category C (5 - 50 kb)			
Manta	95.0/100.0	96.9/100.0	96.3/100.0	96.9/100.0
LinkedSV (short)	66.3/100.0	82.5/100.0	86.3/100.0	86.9/100.0
LinkedSV (linked)	66.3/96.6	66.3/98.2	63.1/97.9	55.0/97.9
LongRanger (short)	20.0/100.0	1.3/100.0	0.0/0.0	0.0/0.0
LongRanger (linked)	26.3/86.4	26.3/82.4	26.3/82.6	27.5/81.5
VALOR2	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
NAIBR	90.0/98.7	88.1/100.0	82.5/100.0	82.5/100.0
LEVIATHAN	31.3/68.3	44.4/73.2	45.0/73.8	43.8/73.5
	SV length category D (50 - 250 kb)			
Manta	95.0/100.0	95.0/100.0	95.0/100.0	95.0/100.0
LinkedSV (short)	25.0/100.0	32.5/100.0	32.5/100.0	30.0/100.0
LinkedSV (linked)	60.0/100.0	65.0/93.7	62.5/100.0	50.0/100.0
LongRanger (short)	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LongRanger (linked)	25.0/63.6	20.0/66.7	20.0/66.7	20.0/75.0
VALOR2	42.5/89.4	40.0/91.3	40.0/96.2	40.0/90.0
NAIBR	97.5/100.0	100.0/100.0	100.0/100.0	95.0/100.0
LEVIATHAN	20.0/52.3	32.5/46.6	37.5/46.8	35.0/45.3
	SV length category E (0.25 - 1 Mb)			
Manta	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LinkedSV (short)	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LinkedSV (linked)	50.0/100.0	50.0/100.0	50.0/100.0	50.0/100.0
LongRanger (short)	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LongRanger (linked)	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
VALOR2	0.0/0.0	0.0/0.0	25.0/100.0	25.0/100.0
NAIBR	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LEVIATHAN	25.0/66.67	50.0/66.67	25.0/50.0	50.0/100.0

Table 3: Sensitivity/precision (%) of structural variant (SV) callers to detect inversions of the SV length categories C (5 - 50 kb), D (50 - 250 kb), and E (0.25 - 1 Mb) in complex simulations.

	Sequencing coverage			
SV caller	45x	90x	135x	180x
	SV length category C (5 - 50 kb)			
Manta	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LinkedSV	34.4/100.0	40.6/100.0	25.0/100.0	18.8/100.0
LongRanger	18.8/100.0	18.8/100.0	18.8/100.0	25.0/100.0
VALOR2	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
NAIBR	56.3/100.0	50.0/100.0	37.5/100.0	31.3/100.0
LEVIATHAN	71.9/90.5	81.3/84.0	81.3/82.8	80.6/81.3
	SV length category D (50 - 250 kb)			
Manta	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LinkedSV	62.5/100.0	68.75/100.0	50.0/100.0	37.5/100.0
LongRanger	25.0/66.7	25.0/100.0	37.5/100.0	37.5/100.0
VALOR2	37.5/24.3	43.8/26.1	37.5/26.7	37.5/25.0
NAIBR	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LEVIATHAN	62.5/100.0	75.0/82.9	87.5/87.5	87.5/82.9
	SV length category E (0.25 - 1 Mb)			
Manta	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LinkedSV	68.8/100.0	75.0/100.0	62.5/100.0	37.5/100.0
LongRanger	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
VALOR2	37.5/43.7	62.5/83.3	62.5/81.7	68.8/84.5
NAIBR	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LEVIATHAN	68.8/100.0	87.5/100.0	87.5/100.0	87.5/100.0

Table 4: Sensitivity/precision (%) of structural variant (SV) callers to detect duplications of the SV length categories C (5 - 50 kb), D (50 - 250 kb), and E (0.25 - 1 Mb) in complex simulations.

	Sequencing coverage			
SV caller	45x	90x	135x	180x
	SV length category C (5 - 50 kb)			
Manta	95.8/100.0	95.8/100.0	95.8/100.0	95.8/100.0
LinkedSV	8.3/100.0	8.3/100.0	0.0/0.0	0.0/0.0
LongRanger	12.5/100.0	8.3/100.0	12.5/100.0	12.5/100.0
NAIBR	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LEVIATHAN	22.9/18.0	37.5/22.0	37.5/23.8	33.3/21.1
	SV length category D (50 - 250 kb)			
Manta	90.0/100.0	95.0/100.0	95.0/100.0	95.0/100.0
LinkedSV	40.0/100.0	50.0/100.0	50.0/100.0	50.0/100.0
LongRanger	40.0/80.0	45.0/77.8	50.0/80.0	60.0/83.3
NAIBR	85.0/100.0	80.0/100.0	80.0/100.0	75.0/100.0
LEVIATHAN	15.0/44.4	30.0/46.2	25.0/37.5	25.0/42.9
	SV length category E (0.25 - 1 Mb)			
Manta	75.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LinkedSV	25.0/100.0	50.0/100.0	50.0/100.0	37.5/100.0
LongRanger	0.0/0.0	25.0/100.0	25.0/100.0	50.0/100.0
NAIBR	100.0/100.0	100.0/100.0	100.0/100.0	100.0/100.0
LEVIATHAN	0.0/0.0	25.0/50.0	25.0/50.0	25.0/50.0

Table 5: Sensitivity/precision (%) of structural variant (SV) callers to detect insertions of the SV length categories A-E (50 bp - 1Mb) in complex simulations.

	Sequencing coverage			
SV caller	45x	90x	135x	180x
	SV length categories A-E (50 bp - 1 Mb)			
Manta	91.6/98.5	94.6/98.2	94.4/97.6	93.4/97.3
Novel-X	18.6/98.3	74.5/98.5	82.1/98.5	86.3/98.5
Novel-X (2 BND)[1]	9.1/47.9	72.2/95.8	81.6/98.0	86.0/98.4

[1]Start and end break points were considered for evaluation

Table 6: Resources used by SV callers in the case of 180x sequencing coverage and in complex simulations. For details see material and methods.

SV caller	Walltime (h)	CPU time (h)	MEM (GB)	VMEM (GB)	Number of CPU used
Manta	00:05:05	00:09:02	0.11	1.83	2
LinkedSV	03:01:44	09:50:53	4.77	12.45	4
LongRanger[1]	-	-	-	-	-
VALOR2	00:10:05	00:09:10	5.35	6.32	1
NAIBR	05:54:32	05:54:08	14.06	15.21	1
LEVIATHAN	09:59:03	19:32:14	32.31	28.23	2
Novel-X	09:14:34	33:11:24	7.39	8.88	4

[1]SV calling during LongRanger wgs mapping

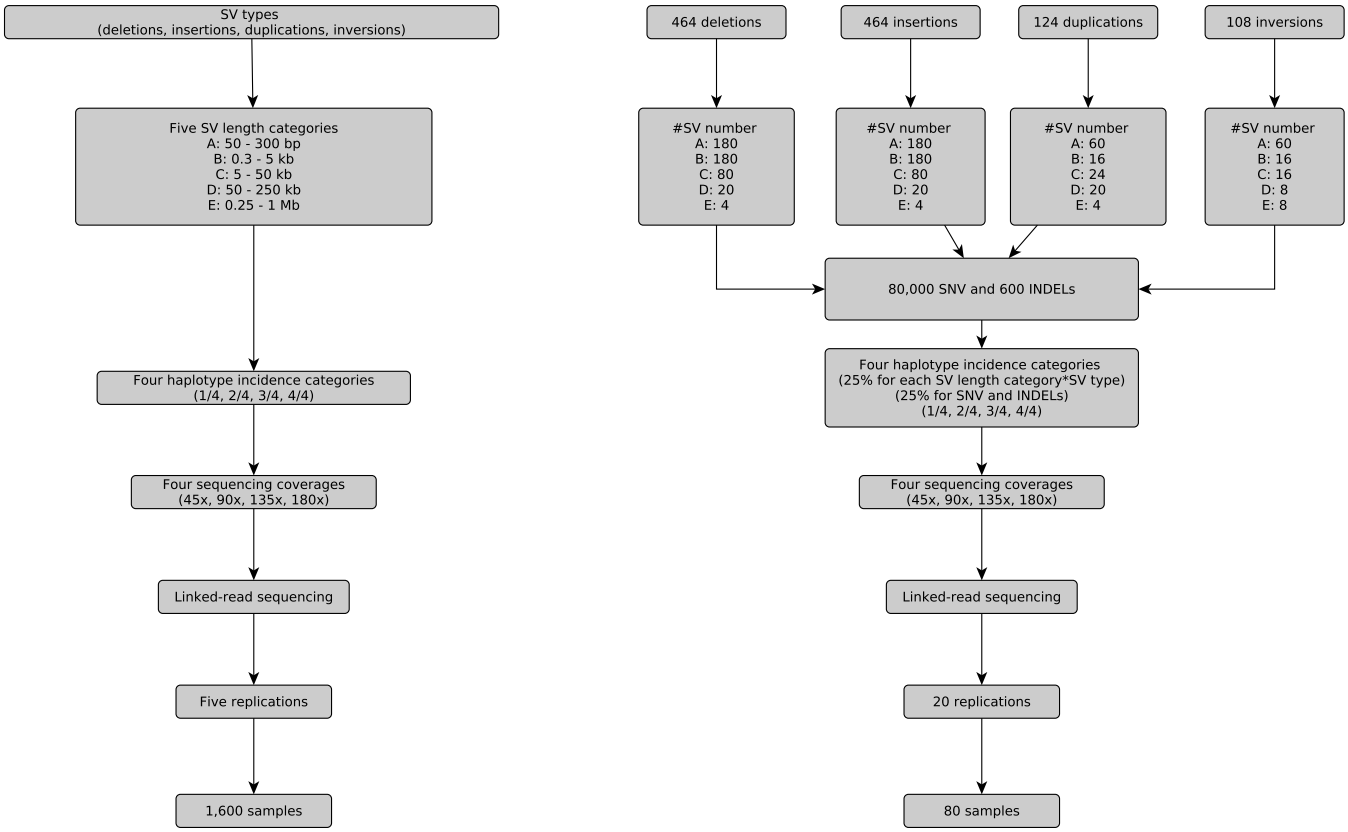

Fig. 1: Overview of the simple (left) and complex (right) simulations.

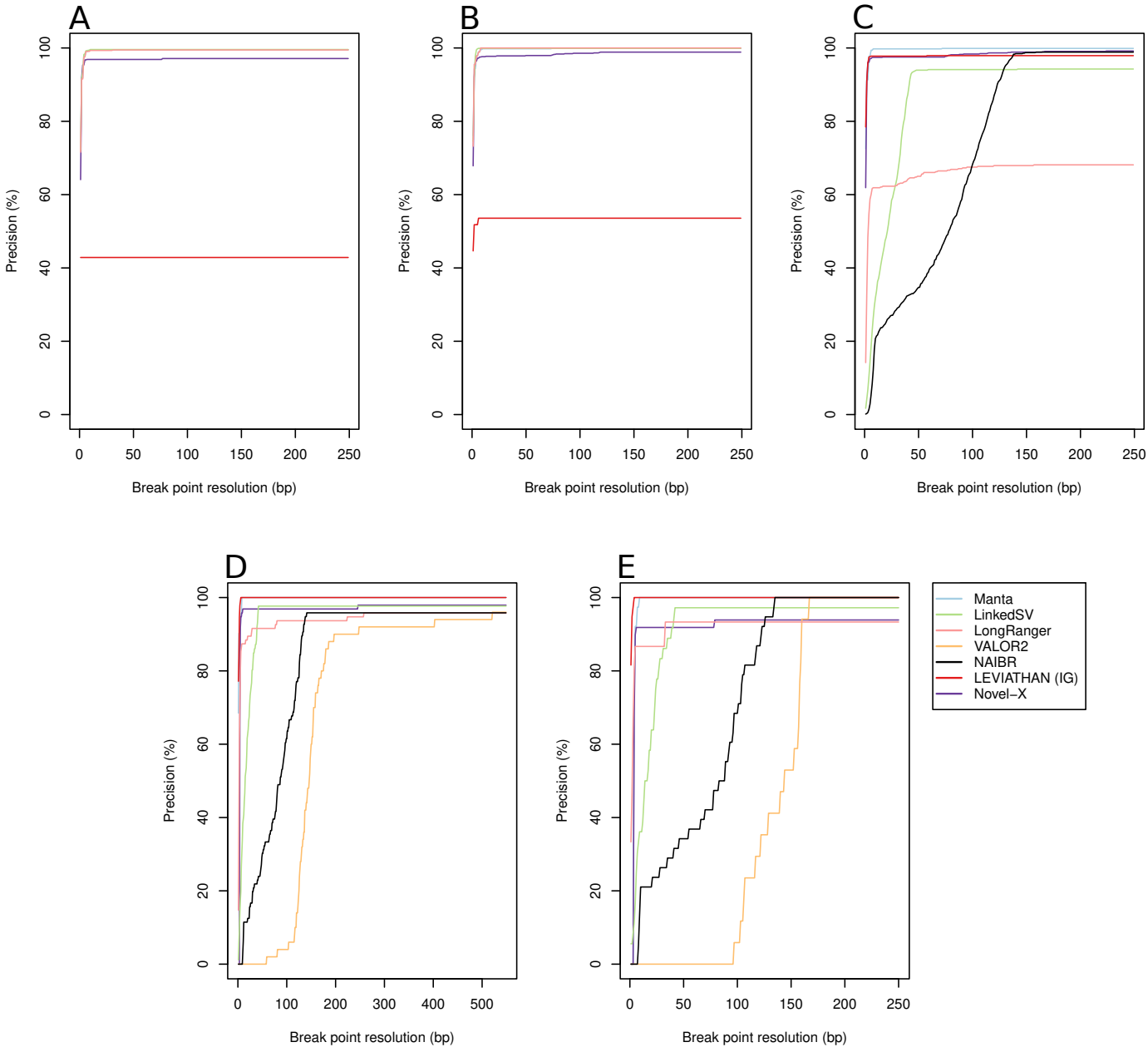

Fig. 2: Break point resolution in bp of the different SV callers for five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) based on the detection of homozygous (4/4) deletions using a linked-read sequencing coverage of 135x.

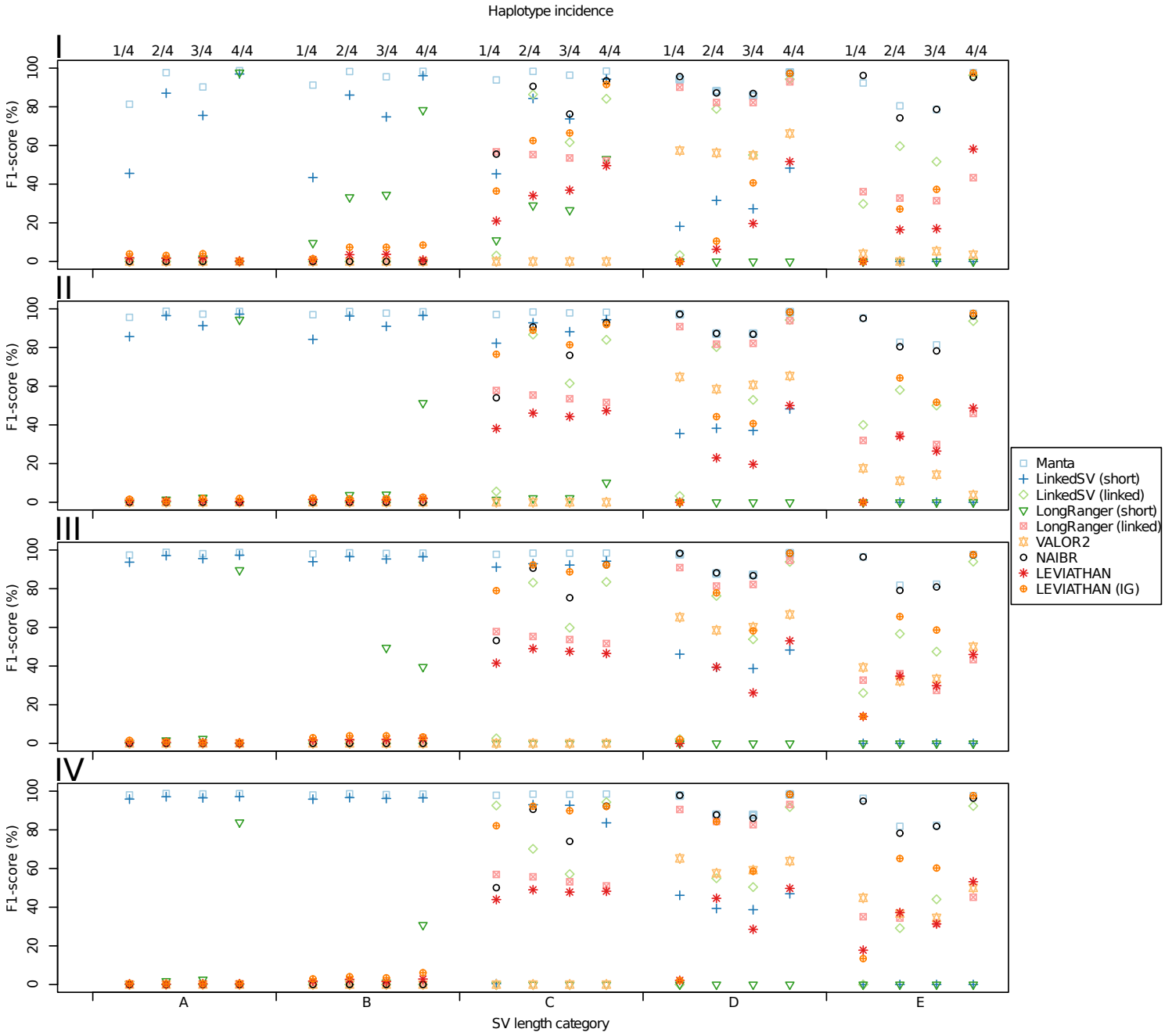

Fig. 3: F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of deletions of five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) and four haplotype incidences (1/4, 2/4, 3/4, 4/4) using different SV callers (for details see Material & Methods) based on 45x (I), 90x (II), 135x (III), and 180x (IV) coverage of linked-read sequencing.

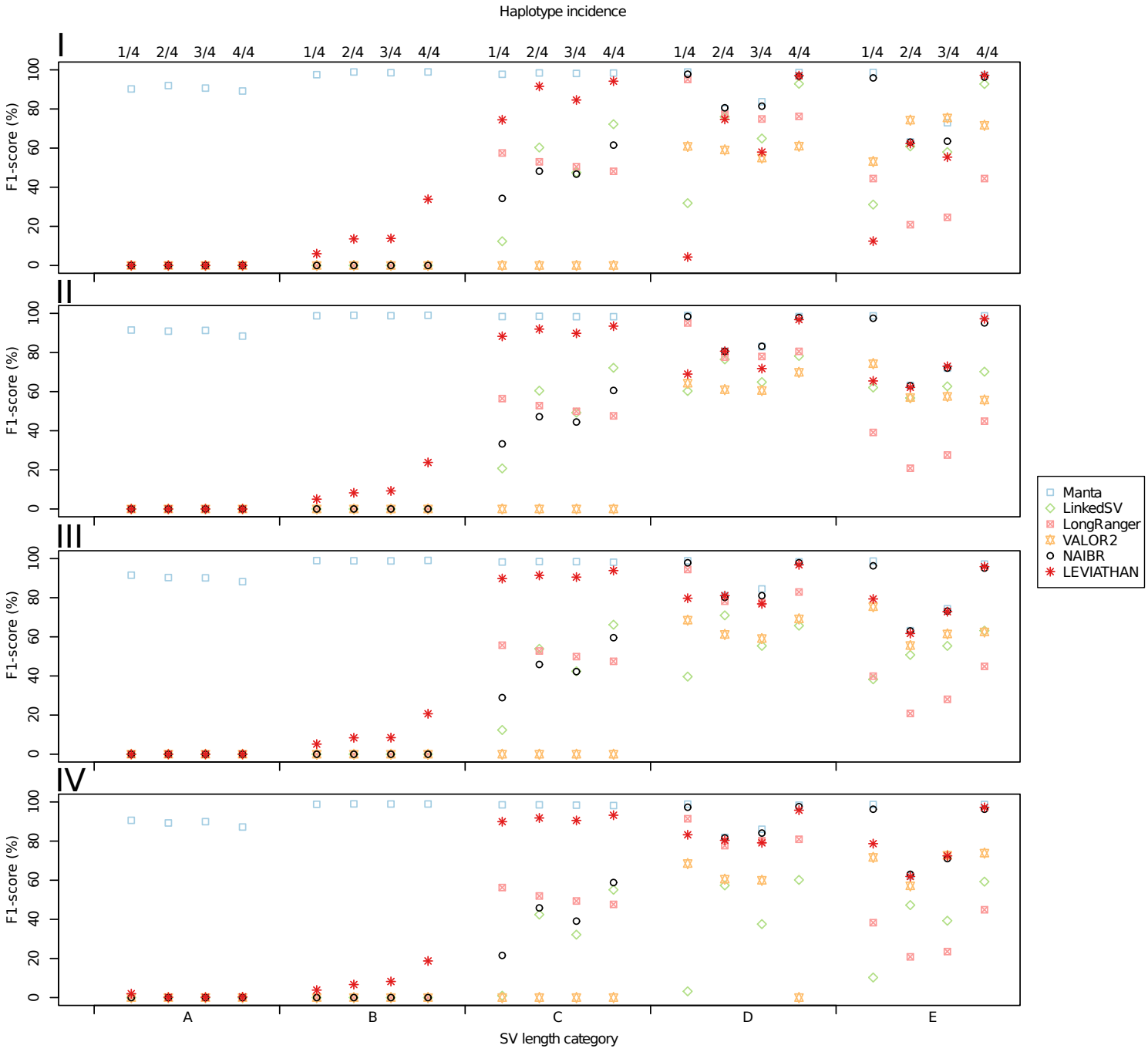

Fig. 4: F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of inversions of five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) and four haplotype incidences (1/4, 2/4, 3/4, 4/4) using different SV callers (for details see Material & Methods) based on 45x (I), 90x (II), 135x (III), and 180x (IV) coverage of linked-read sequencing.

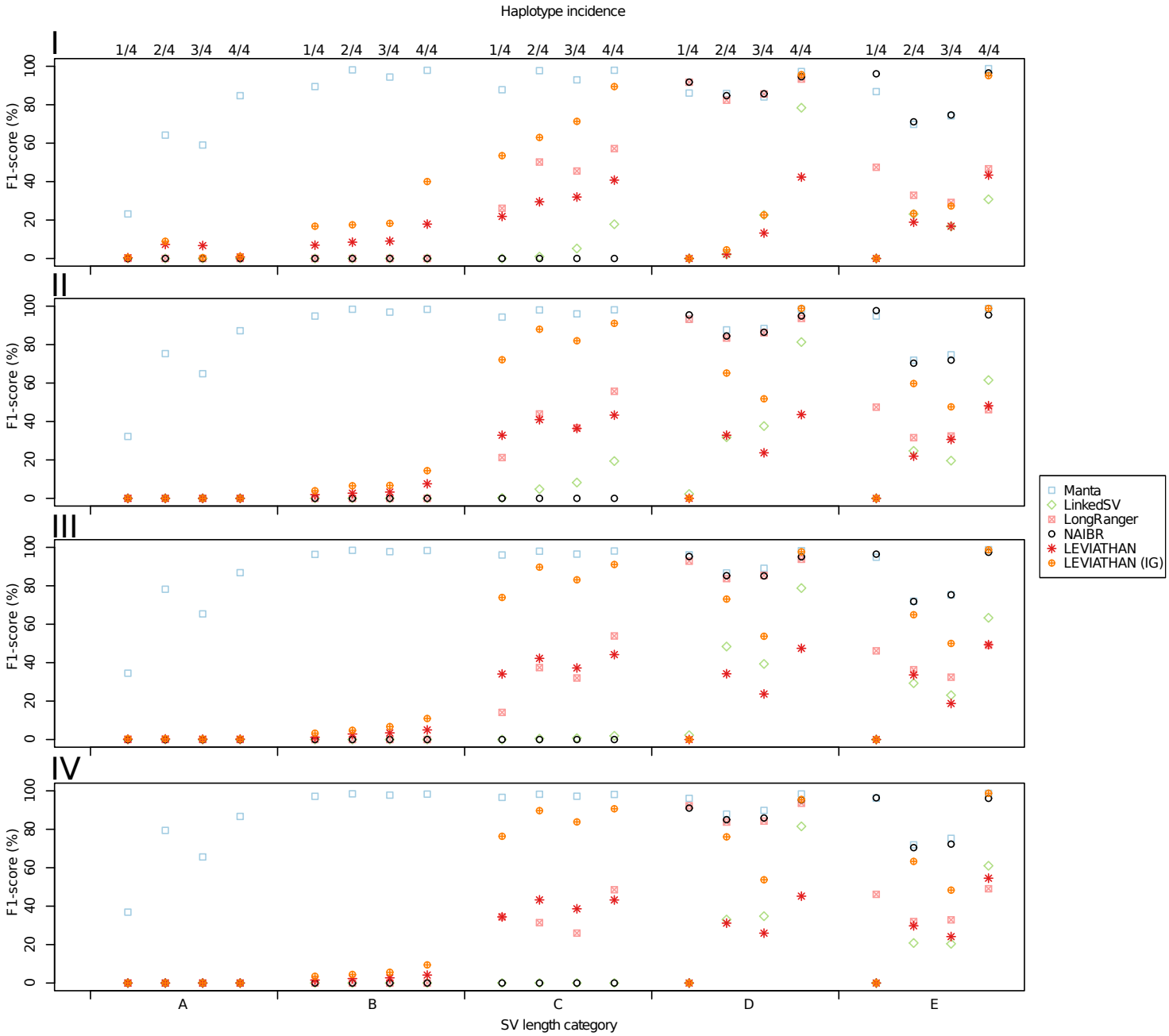

Fig. 5: F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of duplications of five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) and four haplotype incidences (1/4, 2/4, 3/4, 4/4) using different SV callers (for details see Material & Methods) based on 45x (I), 90x (II), 135x (III), and 180x (IV) coverage of linked-read sequencing.

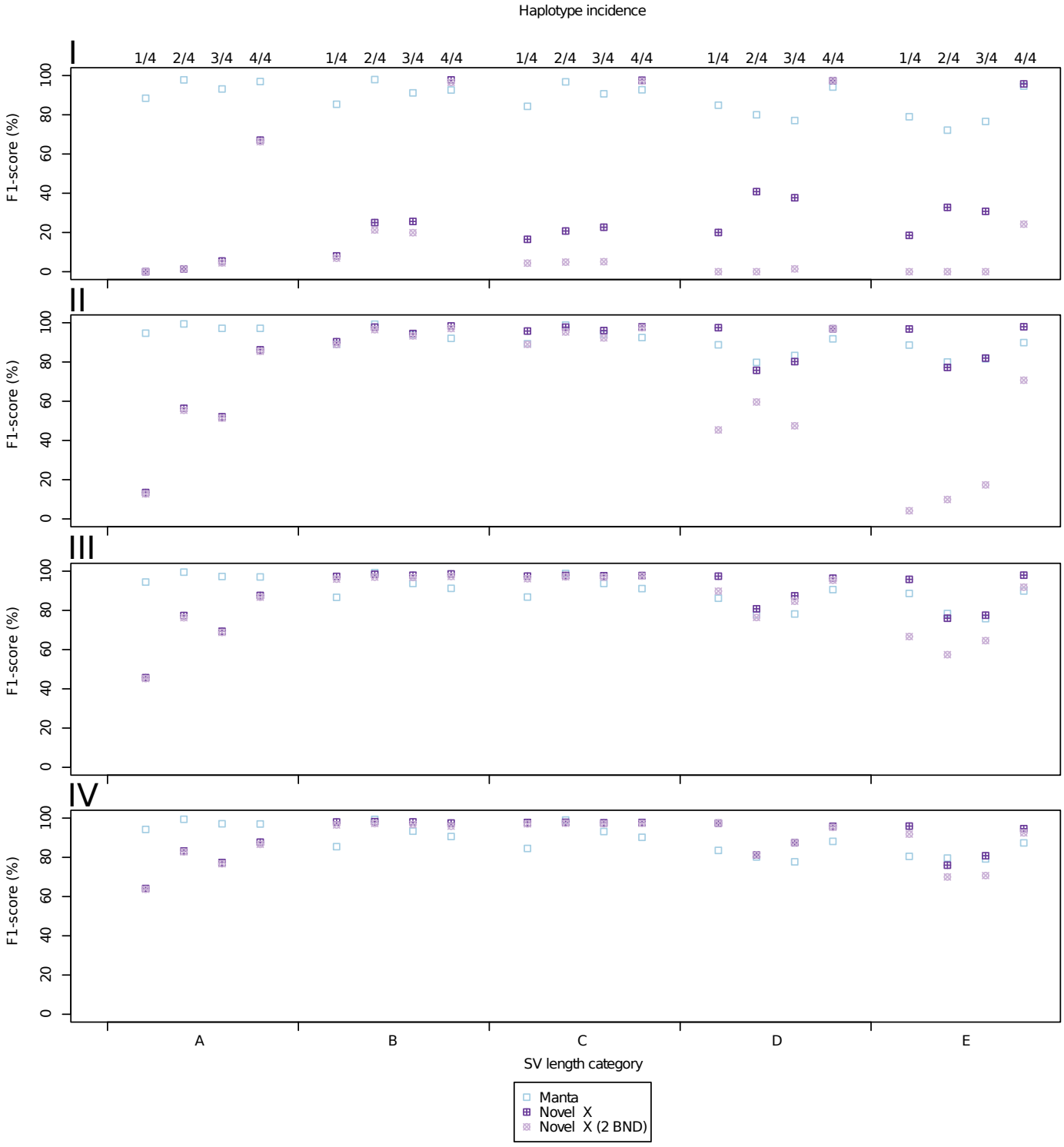

Fig. 6: F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of insertions of five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) and four haplotype incidences (1/4, 2/4, 3/4, 4/4) using different SV callers (for details see Material & Methods) based on 45x (I), 90x (II), 135x (III), and 180x (IV) coverage of linked-read sequencing.

SUPPLEMENTARY INFORMATION

Table S1: Break point resolution in bp allowed to determine the detected structural variants (SV) as true positive SV for the different SV types and SV length categories.

SV type	A (50 - 300 bp)	B (0.3 - 5 kb)	C (5 - 50 kb)	D (50 - 250 kb)	E (0.25 - 1 Mb)
Deletions	$\leq$ 10 bp	$\leq$ 50 bp	$\leq$ 160 bp	$\leq$ 550 bp	$\leq$ 250 bp
Inversions	$\leq$ 10 bp	$\leq$ 50 bp	$\leq$ 160 bp	$\leq$ 800 bp	$\leq$ 550 bp
Duplications	$\leq$ 10 bp	$\leq$ 50 bp	$\leq$ 160 bp	$\leq$ 250 bp	$\leq$ 250 bp
Insertions (2 BND)[1]	$\leq$ 10 bp	$\leq$ 50 bp	$\leq$ 160 bp	$\leq$ 550 bp	$\leq$ 250 bp
Insertions	$\leq$ 10 bp	$\leq$ 10 bp	$\leq$ 10 bp	$\leq$ 10 bp	$\leq$ 10 bp

[1]Start and end break points were considered for evaluation

Table S2: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category A (50 - 300 bp) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	68.6/99.7	95.6/99.7	82.5/99.6	97.6/99.7
LinkedSV (short)	29.6/100.0	77.1/99.6	60.9/99.7	94.3/99.7
LongRanger (short)	0.1/100.0	0.4/69.1	0.8/87.5	95.7/99.7
LEVIATHAN	1.0/52.0	0.9/46.6	1.0/51.0	0.1/100.0
LEVIATHAN (IG)	2.0/100.0	1.6/88.2	2.1/97.1	0.1/100.0
	90x sequence coverage			
Manta	91.9/99.7	98.0/99.6	95.0/99.6	97.9/99.5
LinkedSV (short)	75.2/99.6	93.5/99.7	84.1/99.8	94.9/99.5
LongRanger (short)	0.2/75.0	0.7/84.6	1.2/85.7	89.8/99.6
LEVIATHAN	0.4/38.5	0.3/37.5	0.5/41.7	0.2/50.0
LEVIATHAN (IG)	0.8/80.0	0.2/76.9	1.0/83.3	1.0/83.3
	135x sequence coverage			
Manta	95.0/99.6	98.0/99.6	96.8/99.5	98.0/99.5
LinkedSV (short)	88.4/99.7	94.5/99.7	91.6/99.7	94.9/99.6
LongRanger (short)	0.1/75.0	0.8/80.0	1.2/92.3	81.7/99.5
LEVIATHAN	0.3/27.3	0.3/42.9	0.2/39.6	0.1/33.3
LEVIATHAN (IG)	0.8/72.7	0.4/74.2	0.2/66.7	0.2/66.7
	180x sequence coverage			
Manta	96.6/99.7	98.1/99.6	97.3/99.6	98.0/99.5
LinkedSV (short)	92.5/99.7	94.4/99.6	93.3/99.7	94.9/99.4
LongRanger (short)	0.2/58.3	0.9/83.3	1.3/86.7	72.4/99.3
LEVIATHAN	0.2/41.7	0.1/25.0	0.2/58.3	0.2/33.3
LEVIATHAN (IG)	0.3/66.7	0.2/50.0	0.2/0.0	0.3/50.0

Table S3: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category B (0.3 - 5 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	83.9/100.0	96.8/99.9	91.4/100.0	97.1/99.7
LinkedSV (short)	27.8/100.0	75.5/100.0	59.9/99.8	93.4/98.8
LongRanger (short)	5.1/74.6	20.9/80.7	21.7/83.7	68.2/92.0
LEVIATHAN	0.5/33.3	1.9/37.5	2.0/37.7	0.4/16.7
LEVIATHAN (IG)	0.7/53.9	3.8/78.4	3.8/78.4	4.5/83.0
	90x sequence coverage			
Manta	94.1/100.0	97.7/99.6	95.7/99.8	97.5/99.6
LinkedSV (short)	72.7/100.0	92.8/100.0	83.5/100.0	94.03/99.8
LongRanger (short)	0.2/41.7	1.9/76.9	2.0/80.0	34.74/97.2
LEVIATHAN	0.7/26.9	0.4/12.5	0.6/20.7	1.0/26.5
LEVIATHAN (IG)	1.1/48.3	0.9/31.0	0.9/31.0	1.3/40.5
	135x sequence coverage			
Manta	96.3/99.9	97.7/99.6	96.8/99.7	97.4/99.5
LinkedSV (short)	88.6/100.0	93.3/100.0	91.1/100.0	93.9/99.8
LongRanger (short)	0.1/100.0	0.3/75.0	47.4/99.0	24.8/98.7
LEVIATHAN	0.9/27.3	1.0/25.6	1.1/28.2	1.4/27.2
LEVIATHAN (IG)	1.5/50.0	2.1/48.7	2.1/48.7	1.7/47.2
	180x sequence coverage			
Manta	96.2/99.9	97.7/99.6	97.0/99.6	97.5/99.6
LinkedSV (short)	92.0/100.0	93.5/100.0	92.7/99.9	93.8/99.6
LongRanger (short)	0.1/50.0	0.4/90.0	0.5/92.9	18.2/98.8
LEVIATHAN	0.9/28.1	1.4/27.3	0.8/20.0	1.5/22.9
LEVIATHAN (IG)	1.5/43.9	2.1/46.0	1.8/43.6	3.3/48.5

Table S4: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category C (5 - 50 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	88.4/100.0	96.7/100.0	92.9/100.0	97.0/100.0
LinkedSV (short)	29.3/100.0	73.0/99.8	58.5/99.8	89.3/99.9
LinkedSV (linked)	1.6/100.0	77.1/98.1	45.2/97.0	78.5/90.6
LongRanger (short)	5.8/98.3	17.0/99.3	15.3/99.3	36.1/100.0
LongRanger (linked)	40.8/91.8	42.3/79.1	41.7/73.3	42.0/68.6
NAIBR	38.7/98.3	83.9/97.0	62.6/97.0	88.2/99.1
LEVIATHAN	12.9/56.9	24.9/54.1	27.6/57.0	46.6/52.9
LEVIATHAN (IG)	22.3/99.2	45.6/99.2	49.9/99.2	85.2/98.6
	90x sequence coverage			
Manta	94.3/100.0	96.8/100.0	95.9/100.0	96.6/100.0
LinkedSV (short)	69.9/100.0	86.6/99.9	79.0/99.9	89.3/99.9
LinkedSV (linked)	2.9/100.0	78.3/97.5	45.0/97.3	76.4/93.9
LongRanger (short)	0.7/95.5	1.1/100.0	1.1/100.0	5.4/95.7
LongRanger (linked)	41.8/93.9	42.5/79.9	42.6/75.1	42.0/67.3
NAIBR	37.1/99.1	84.4/97.4	62.4/97.2	87.2/98.7
LEVIATHAN	31.1/52.8	41.2/52.5	37.8/53.1	44.2/50.8
LEVIATHAN (IG)	62.4/99.1	80.8/98.9	69.5/98.4	86.0/98.6
	135x sequence coverage			
Manta	95.6/100.0	96.8/100.0	96.8/100.0	97.0/100.0
LinkedSV (short)	83.8/100.0	87.0/99.9	85.7/99.9	89.1/99.9
LinkedSV (linked)	1.3/100.0	72.2/99.0	43.1/98.3	75.1/94.3
LongRanger (short)	0.1/100.0	0.1/100.0	0.1/50.0	0.2/66.7
LongRanger (linked)	41.6/94.7	42.6/80.0	42.6/74.0	41.7/67.8
NAIBR	36.4/99.1	84.7/97.2	61.6/97.3	86.5/98.8
LEVIATHAN	35.1/52.1	45.8/52.7	43.0/53.1	43.8/49.2
LEVIATHAN (IG)	65.8/98.8	86.4/98.8	80.4/98.9	86.7/98.2
	180x sequence coverage			
Manta	95.8/100.0	97.0/100.0	96.6/100.0	97.2/100.0
LinkedSV (short)	0.2/83.3	87.2/99.9	86.5/99.9	74.1/95.1
LinkedSV (linked)	86.5/99.9	54.6/99.2	40.2/98.6	89.3/99.9
LongRanger (short)	0.0/0.0	0.1/50.0	0.1/100.0	0.2/50.0
LongRanger (linked)	40.5/96.0	42.3/81.5	41.6/73.6	41.5/68.2
NAIBR	33.5/99.3	85.1/97.2	59.5/97.0	86.2/98.9
LEVIATHAN	37.7/52.6	46.1/52.5	44.0/52.5	45.7/50.7
LEVIATHAN (IG)	70.2/98.8	86.1/98.9	82.1/99.1	86.5/97.9

Table S5: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category D (50 - 250 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	89.0/100.0	91.2/85.0	87.8/83.7	95.6/100.0
LinkedSV (short)	10.0/100.0	19.4/89.5	16.7/76.9	31.8/100.0
LinkedSV (linked)	1.7/100.0	65.9/92.7	40.9/89.7	90.1/98.7
LongRanger (linked)	91.2/87.0	89.3/73.3	89.3/74.1	94.4/91.4
VALOR2	42.9/86.7	44.1/77.4	43.3/75.6	50.5/94.1
NAIBR	95.6/95.6	93.4/81.7	92.5/81.9	97.8/96.7
LEVIATHAN	0.0/0.0	3.3/60.0	12.2/50.0	50.5/52.8
LEVIATHAN (IG)	0.0/0.0	5.6/100.0	26.4/90.0	95.6/100
	90x sequence coverage			
Manta	94.5/100.0	92.5/84.3	91.4/83.3	96.7/100.0
LinkedSV (short)	21.6/100.0	24.4/84.9	24.7/84.9	31.8/100.0
LinkedSV (linked)	1.7/100.0	69.3/93.0	38.7/83.7	90.1/98.8
LongRanger (linked)	93.4/88.4	88.9/72.2	89.3/74.1	94.4/93.3
VALOR2	49.5/93.9	46.2/81.1	47.3/84.6	50.5/92.2
NAIBR	97.8/97.8	93.4/82.2	92.5/81.9	97.8/98.9
LEVIATHAN	0.0/0.0	15.4/45.2	12.2/50.0	50.0/50.0
LEVIATHAN (IG)	0.0/0.0	29.7/91.7	26.4/90.0	96.7/98.9
	135x sequence coverage			
Manta	95.6/100.0	92.5/84.3	91.4/83.3	96.7/100.0
LinkedSV (short)	30.0/100.0	25.8/82.8	25.8/85.7	31.8/100.0
LinkedSV (linked)	1.1/100.0	65.6/91.0	38.9/87.5	88.6/98.6
LongRanger (linked)	92.2/90.4	89.3/72.7	89.3/75.7	95.6/93.7
VALOR2	49.5/95.7	46.2/79.6	47.3/83.3	51.6/94.0
NAIBR	97.8/98.9	93.4/83.0	93.4/82.5	98.4/98.3
LEVIATHAN	0.0/0.0	33.0/50.9	19.7/39.8	52.8/53.3
LEVIATHAN (IG)	1.1/25.0	64.4/93.9	44.1/86.2	96.7/100.0
	180x sequence coverage			
Manta	95.6/100.0	92.5/84.3	92.5/83.7	96.7/100.0
LinkedSV (short)	30.0/100.0	25.8/82.8	25.8/85.7	30.7/100.0
LinkedSV (linked)	0.0/0.0	38.6/93.3	34.4/89.2	85.2/98.7
LongRanger (linked)	94.5/86.9	90.0/75.4	89.3/76.4	92.2/92.2
VALOR2	49.5/95.7	46.2/80.8	46.2/81.1	49.5/87.8
NAIBR	96.7/97.9	92.5/80.7	90.1/83.0	97.8/98.9
LEVIATHAN	1.1/100.0	38.9/50.0	21.1/44.2	48.9/50.6
LEVIATHAN (IG)	1.1/100.0	72.7/92.3	43.3/84.8	97.8/100.0

Table S6: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category E (250 - 1 Mb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	85.7/100.0	82.5/76.1	82.5/73.3	95.5/100.0
LinkedSV (linked)	17.5/100.0	43.6/88.0	40.0/76.2	93.18/100.0
LongRanger (linked)	29.6/46.2	25.0/41.9	27.5/36.7	29.55/92.3
VALOR2	2.6/9.1	0.0/0.0	3.5/10.7	2.5/5.9
NAIBR	95.5/95.2	90.0/63.2	87.5/71.4	95.5/95.1
LEVIATHAN	0.0/0.0	10.2/53.1	11.4/44.4	56.8/59.5
LEVIATHAN (IG)	0.0/0.0	18.2/100.0	25.0/83.3	95.0/100.0
	90x sequence coverage			
Manta	90.9/100.0	87.5/75.0	87.5/74.5	95.5/100.0
LinkedSV (linked)	25.0/100.0	46.2/78.3	38.6/70.0	90.9/97.5
LongRanger (linked)	27.3/38.7	29.6/41.4	25.0/37.0	31.8/92.9
VALOR2	11.4/39.4	7.5/23.1	9.5/28.6	2.6/8.3
NAIBR	95.5/90.7	90.0/69.8	90.0/69.2	95.5/97.4
LEVIATHAN	0.0/0.0	29.6/40.5	20.5/37.5	48.7/48.7
LEVIATHAN (IG)	0.0/0.0	50.0/77.8	38.6/70.8	95.5/100.0
	135x sequence coverage			
Manta	93.2/100.0	87.5/73.5	87.5/75.0	95.5/100.0
LinkedSV (linked)	15.0/100.0	43.6/81.0	35.0/73.7	88.6/100.0
LongRanger (linked)	28.4/39.0	27.5/48.0	25.0/30.3	29.6/92.3
VALOR2	26.2/90.0	20.5/75.0	21.4/75.0	33.3/90.9
NAIBR	95.2/97.6	90.0/69.2	88.6/73.5	95.5/97.7
LEVIATHAN	7.5/100.0	27.3/38.6	22.7/38.5	45.5/46.5
LEVIATHAN (IG)	7.5/100.0	52.3/92.0	45.0/78.3	95.5/100.0
	180x sequence coverage			
Manta	92.9/100.0	87.5/73.5	87.5/75.0	95.5/100.0
LinkedSV (linked)	0.0/0.0	17.5/87.5	29.6/78.6	85.7/100.0
LongRanger (linked)	28.6/41.7	27.5/45.8	27.3/33.3	31.8/86.7
VALOR2	29.6/100.0	25.0/69.2	22.7/71.4	33.3/100.0
NAIBR	94.9/95.1	90.0/67.9	90.0/72.6	95.5/97.4
LEVIATHAN	10.0/80.0	31.8/48.2	28.2/33.3	52.4/53.9
LEVIATHAN (IG)	7.4/100.0	57.5/88.9	50.0/73.1	95.6/100.0

Table S7: Sensitivity/precision (%) of Manta to detect inversions of the structural variant length category A (50 - 300 bp) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	86.9/93.9	93.7/90.3	90.3/91.4	91.1/86.9
	90x sequence coverage			
Manta	91.7/91.1	94.2/87.7	94.1/88.6	92.4/85.1
	135x sequence coverage			
Manta	93.3/89.7	94.6/86.4	93.1/86.9	93.0/83.3
	180x sequence coverage			
Manta	94.0/88.0	94.3/84.8	93.9/86.0	93.4/82.6

Table S8: Sensitivity/precision (%) of structural variant (SV) callers to detect inversions of the SV length category B (0.3 - 5 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	95.4/99.8	98.2/99.6	97.3/99.8	98.5/99.2
LEVIATHAN	3.16/56.1	7.4/71.4	7.6/77.7	21.1/86.1
	90x sequence coverage			
Manta	97.8/99.8	98.4/99.6	98.1/99.4	98.4/99.5
LEVIATHAN	2.7/42.2	4.4/55.4	5.1/58.8	14.2/75.9
	135x sequence coverage			
Manta	98.2/99.7	98.3/99.5	98.1/99.6	98.4/99.4
LEVIATHAN	2.8/47.2	4.5/55.7	4.6/52.4	12.1/71.0
	180x sequence coverage			
Manta	98.0/99.7	98.5/99.4	98.3/99.5	98.4/99.3
LEVIATHAN	2.1/32.4	3.6/44.8	4.5/50.0	11.2/64.2

Table S9: Sensitivity/precision (%) of structural variant (SV) callers to detect inversions of the SV length category C (5 - 50 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	95.6/99.9	97.6/99.4	96.8/99.6	97.5/99.4
LinkedSV	6.6/98.4	43.4/98.6	31.2/98.9	57.0/98.6
LongRanger	40.9/95.9	41.5/73.4	41.1/66.2	42.0/56.0
NAIBR	20.9/96.3	32.3/96.3	30.9/97.2	45.3/96.3
LEVIATHAN	60.0/98.4	86.2/98.0	74.3/98.3	90.8/97.9
	90x sequence coverage			
Manta	97.4/99.4	97.5/99.5	97.1/99.6	97.3/99.5
LinkedSV	11.6/100.0	43.5/99.2	32.7/99.3	56.7/99.2
LongRanger	40.0/95.9	41.7/72.0	40.7/64.1	41.7/54.9
NAIBR	20.0/97.4	31.1/98.0	28.9/97.5	43.9/98.0
LEVIATHAN	80.4/97.8	87.0/98.0	83.5/97.7	90.3/97.0
	135x sequence coverage			
Manta	97.0/99.6	97.4/99.5	97.5/99.4	97.5/99.3
LinkedSV	6.6/100.0	36.9/99.4	26.9/99.6	49.8/98.8
LongRanger	39.5/95.0	41.7/70.2	41.3/63.7	41.9/54.4
NAIBR	16.9/98.7	30.0/98.1	26.9/98.3	42.9/98.3
LEVIATHAN	82.9/97.9	87.0/96.7	85.3/97.0	90.3/96.7
	180x sequence coverage			
Manta	97.6/99.5	97.4/99.8	97.5/99.5	97.2/98.9
LinkedSV	0.5/100.0	26.9/99.5	19.2/98.8	38.2/99.1
LongRanger	39.8/95.4	41.5/69.4	40.7/63.1	42.2/54.0
NAIBR	12.1/99.0	29.9/98.4	24.4/98.7	41.9/98.0
LEVIATHAN	84.0/97.1	87.0/97.0	85.2/97.4	90.4/95.9

Table S10: Sensitivity/precision (%) of structural variant (SV) callers to detect inversions of the SV length category D (50 - 250 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	97.9/100.0	88.3/76.0	85.9/83.7	97.8/98.9
LinkedSV	19.2/100.0	70.7/82.3	52.3/84.5	87.8/98.8
LongRanger	93.6/95.6	83.7/67.2	82.6/68.5	91.9/65.1
VALOR2	45.2/94.0	47.3/78.6	43.2/73.1	48.9/81.1
NAIBR	97.9/97.8	90.4/76.0	85.9/77.5	97.8/96.7
LEVIATHAN	2.2/87.5	64.1/89.3	43.5/79.6	98.4/96.8
	90x sequence coverage			
Manta	97.9/100.0	89.4/76.0	85.9/81.3	97.8/98.9
LinkedSV	43.5/100.0	70.7/83.5	51.6/82.5	64.9/98.5
LongRanger	92.6/97.8	84.0/69.3	82.6/73.8	91.4/70.7
VALOR2	48.9/95.7	50.5/78.7	47.9/79.0	55.4/92.7
NAIBR	97.9/98.9	90.4/76.0	85.9/80.6	97.9/96.8
LEVIATHAN	53.8/97.6	79.3/80.4	63.6/80.0	97.9/94.8
	135x sequence coverage			
Manta	97.9/100.0	89.4/76.7	85.9/83.2	96.8/97.9
LinkedSV	24.7/100.0	60.9/85.1	40.9/84.2	50.0/100.0
LongRanger	90.4/98.8	86.2/69.3	82.6/73.1	92.4/75.2
VALOR2	53.3/93.8	49.5/72.1	49.5/77.6	56.0/90.7
NAIBR	97.8/97.8	90.4/75.2	85.9/78.6	97.9/96.7
LEVIATHAN	67.0/98.3	83.7/78.6	75.0/77.5	97.9/95.6
	180x sequence coverage			
Manta	97.9/100.0	90.4/78.2	88.6/83.7	96.8/98.9
LinkedSV	1.6/100.0	44.6/90.9	23.9/85.6	43.0/100.0
LongRanger	85.1/98.7	85.1/70.2	84.9/77.4	92.4/72.6
VALOR2	53.3/95.8	50.9/70.9	47.0/80.7	0.0/0.0
NAIBR	96.8/98.8	90.4/76.9	86.5/81.7	97.9/97.3
LEVIATHAN	72.7/98.3	83.0/78.0	76.8/81.6	97.9/94.6

Table S11: Sensitivity/precision (%) of structural variant (SV) callers to detect inversions of the SV length category E (0.25 - 1 Mb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	97.4/100.0	67.6/63.9	73.0/71.7	95.0/100.0
LinkedSV	18.4/100.0	54.1/67.7	48.8/69.0	86.5/100.0
LongRanger	29.7/91.7	16.2/43.3	18.9/50.0	29.7/91.7
VALOR2	43.2/65.5	63.4/96.2	63.4/93.1	60.0/92.9
NAIBR	97.5/94.9	67.6/60.0	73.0/62.2	97.5/94.9
LEVIATHAN	6.7/100.0	60.0/63.2	48.7/62.2	97.5/100.0
	90x sequence coverage			
Manta	97.5/100.0	67.6/63.9	73.0/71.1	97.5/100.0
LinkedSV	45.0/100.0	54.1/65.5	55.3/71.4	54.1/100.0
LongRanger	24.3/92.3	16.2/43.3	18.9/47.1	29.7/92.3
VALOR2	63.4/96.2	50.7/65.6	51.4/63.6	50.6/61.9
NAIBR	97.5/97.4	67.6/61.5	73.0/67.4	97.5/94.9
LEVIATHAN	48.7/100.0	62.2/62.2	73.0/68.9	97.5/100.0
	135x sequence coverage			
Manta	97.5/100.0	67.6/63.9	77.7/71.4	95.0/100.0
LinkedSV	23.7/100.0	46.0/61.5	46.1/69.2	48.7/100.0
LongRanger	25.7/95.8	16.2/43.3	19.2/50.0	29.7/92.9
VALOR2	63.4/93.1	50.0/62.5	54.6/70.3	54.1/73.5
NAIBR	97.5/95.1	67.6/61.5	77.7/69.5	97.5/92.9
LEVIATHAN	65.8/100.0	65.9/60.5	74.3/69.4	97.5/97.2
	180x sequence coverage			
Manta	97.5/100.0	67.6/62.2	73.0/69.6	97.5/100.0
LinkedSV	5.4/100.0	35.1/72.2	27.0/68.8	43.2/95.0
LongRanger	23.7/100.0	13.2/42.4	18.9/43.8	29.7/91.7
VALOR2	60.0/92.9	47.4/72.0	62.0/86.2	63.2/86.2
NAIBR	97.3/100.0	67.6/62.2	73.0/64.7	97.5/97.2
LEVIATHAN	64.9/100.0	67.6/60.5	76.5/69.1	97.3/97.4

Table S12: Sensitivity/precision (%) of Manta to detect duplications of the structural variant length category A (50 - 300 bp) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	13.4/87.1	50.0/89.6	43.5/92.3	80.9/89.0
	90x sequence coverage			
Manta	20.3/77.0	66.7/86.6	51.7/87.0	86.7/87.8
	135x sequence coverage			
Manta	22.7/71.3	71.7/86.0	53.2/84.7	86.8/86.6
	180x sequence coverage			
Manta	24.8/71.3	74.2/85.4	54.2/83.0	87.4/85.9

Table S13: Sensitivity/precision (%) of structural variant (SV) callers to detect duplications of the SV length category B (0.3 - 5 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	80.9/100.0	96.3/99.8	89.4/100.0	96.0/99.8
LEVIATHAN	3.8/39.2	4.7/43.4	5.1/42.7	11.4/41.3
LEVIATHAN (IG)	9.2/94.2	9.7/89.6	10.2/86.3	25.5/92.9
	90x sequence coverage			
Manta	90.3/100.0	96.8/99.9	94.0/99.9	97.2/99.7
LEVIATHAN	1.0/25.0	1.4/24.1	1.7/32.2	4.2/39.4
LEVIATHAN (IG)	2.0/56.1	3.5/58.6	3.6/60.4	9.0/75.0
	135x sequence coverage			
Manta	92.9/100.0	97.2/99.8	95.7/99.9	97.2/99.6
LEVIATHAN	0.6/15.6	1.5/24.6	1.8/33.3	2.7/29.7
LEVIATHAN (IG)	1.7/40.5	2.6/46.9	3.6/58.3	6.0/64.8
	180x sequence coverage			
Manta	94.5/100.0	97.2/99.8	95.8/99.9	97.0/99.7
LEVIATHAN	0.8/18.2	1.2/23.5	1.4/25.9	2.3/26.6
LEVIATHAN (IG)	1.8/41.2	2.4/46.9	3.0/53.7	5.1/59.5

Table S14: Sensitivity/precision (%) of structural variant (SV) callers to detect duplications of the SV length category C (5 - 50 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	78.3/100.0	95.8/100.0	86.8/100.0	96.1/99.9
LinkedSV	0.0/0.0	0.5/100.0	2.7/100.0	9.8/98.8
LongRanger	15.1/96.7	34.1/96.8	29.9/95.6	40.3/96.8
NAIBR	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LEVIATHAN	15.3/40.5	21.5/45.7	25.6/44.0	37.4/44.3
LEVIATHAN (IG)	36.8/98.1	46.3/98.3	56.4/98.6	81.9/98.6
	90x sequence coverage			
Manta	89.3/100.0	96.4/100.0	92.5/100.0	96.6/99.9
LinkedSV	0.1/100.0	2.4/100.0	4.3/100.0	10.8/98.8
LongRanger	12.0/96.1	28.4/97.2	22.7/97.1	39.0/97.4
NAIBR	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LEVIATHAN	26.1/43.7	37.4/45.8	31.2/43.7	40.2/46.6
LEVIATHAN (IG)	56.8/98.8	79.3/98.8	70.1/98.5	84.9/98.5
	135x sequence coverage			
Manta	92.4/100.0	96.4/99.9	93.5/100.0	96.6/99.9
LinkedSV	0.0/0.0	0.2/100.0	0.4/100.0	1.0/91.7
LongRanger	7.6/95.6	23.2/97.2	19.2/96.6	37.3/97.2
NAIBR	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LEVIATHAN	28.3/43.8	39.1/46.4	32.3/44.7	40.9/48.0
LEVIATHAN (IG)	59.5/98.0	82.5/98.5	71.9/98.5	85.1/98.4
	180x sequence coverage			
Manta	93.44/100.0	96.6/99.9	94.6/100.0	96.7/99.6
LinkedSV	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LongRanger	28.6/45.1	18.7/97.5	15.0/96.1	32.6/96.8
NAIBR	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
LEVIATHAN	28.6/45.1	39.8/47.3	33.8/45.3	40.3/46.9
LEVIATHAN (IG)	62.7/98.2	83.2/98.3	73.0/98.1	84.8/98.3

Table S15: Sensitivity/precision (%) of structural variant (SV) callers to detect duplications of the SV length category D (50 - 250 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	75.6/100.0	90.3/81.4	85.0/83.7	94.7/100.0
LinkedSV	0.0/0.0	1.1/86.7	12.9/87.0	64.5/96.7
LongRanger	88.5/95.4	90.4/75.7	90.0/80.4	93.6/93.3
NAIBR	88.9/96.2	91.5/78.9	90.8/82.4	92.6/96.7
LEVIATHAN	0.0/0.0	1.2/50.0	7.8/40.0	40.4/44.4
LEVIATHAN (IG)	0.0/0.0	2.3/100.0	13.3/90.0	93.1/97.8
	90x sequence coverage			
Manta	87.4/100.0	92.2/81.7	90.3/85.6	96.8/100.0
LinkedSV	1.2/100.0	19.2/92.9	24.4/88.0	70.1/95.7
LongRanger	91.1/95.4	90.3/76.4	90.3/82.4	93.6/94.3
NAIBR	93.3/97.7	92.6/77.7	91.5/81.9	93.1/96.6
LEVIATHAN	0.0/0.0	24.5/45.3	17.2/37.8	43.3/43.8
LEVIATHAN (IG)	0.0/0.0	48.9/97.2	38.3/85.7	97.8/100.0
	135x sequence coverage			
Manta	92.6/100.0	93.6/80.7	91.1/85.4	96.8/100.0
LinkedSV	1.1/100.0	31.9/96.8	24.7/89.3	66.2/96.1
LongRanger	90.0/96.6	91.1/77.3	88.2/82.8	93.6/95.4
NAIBR	91.1/98.8	91.5/78.6	88.9/82.2	92.6/97.8
LEVIATHAN	0.0/0.0	28.9/44.4	17.8/35.6	46.7/48.3
LEVIATHAN (IG)	0.0/0.0	61.3/96.2	39.4/84.4	96.6/98.8
	180x sequence coverage			
Manta	92.5/100	93.3/80.5	93.3/85.6	96.8/100.0
LinkedSV	0.0/0.0	20.0/94.7	22.2/95.8	68.8/98.5
LongRanger	88.3/96.5	88.9/77.3	88.2/81.1	92.6/94.6
NAIBR	84.4/98.7	90.4/80.2	87.2/84.0	92.2/97.6
LEVIATHAN	0.0/0.0	26.6/38.2	19.5/38.6	44.4/46.0
LEVIATHAN (IG)	0.0/0.0	62.1/95.2	38.7/86.1	94.4/96.7

Table S16: Sensitivity/precision (%) of structural variant (SV) callers to detect duplications of the SV length category B (0.25 - 1 Mb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	76.7/100.0	73.2/66.7	73.7/73.1	97.6/100.0
LinkedSV	0.0/0.0	13.6/75.0	9.8/57.1	18.2/100.0
LongRanger	31.7/91.7	27.3/41.4	27.1/33.3	31.8/92.3
NAIBR	95.0/97.4	78.1/66.0	79.9/70.7	97.7/95.5
LEVIATHAN	0.0/0.0	11.6/28.2	10.4/33.6	42.5/46.0
LEVIATHAN (IG)	0.0/0.0	14.2/66.7	17.3/68.6	90.7/100.0
	90x sequence coverage			
Manta	90.2/100.0	78.1/69.2	75.6/73.5	97.6/100.0
LinkedSV	0.0/0.0	15.1/64.9	11.6/75.0	45.5/100.0
LongRanger	31.8/100.0	27.3/37.5	26.8/40.0	31.8/92.3
NAIBR	97.7/97.7	78.1/65.1	78.1/68.9	97.7/94.9
LEVIATHAN	0.0/0.0	22.5/21.4	22.6/40.0	48.8/47.7
LEVIATHAN (IG)	0.0/0.0	53.5/72.0	35.9/70.0	100.0/100.0
	135x sequence coverage			
Manta	90.2/100.0	77.0/68.0	78.1/72.7	97.6/100.0
LinkedSV	0.0/0.0	18.3/78.0	14.0/66.7	47.5/100.0
LongRanger	30.8/93.8	27.7/50.0	26.8/40.0	33.3/92.9
NAIBR	95.4/97.6	78./66.6	78.1/72.1	97.4/97.6
LEVIATHAN	0.0/0.0	32.1/34.8	14.0/27.5	48.7/50.0
LEVIATHAN (IG)	0.0/0.0	62.9/73.0	37.2/68.8	97.7/97.6
	180x sequence coverage			
Manta	92.7/100.0	78.1/67.4	78.1/72.7	97.4/100.0
LinkedSV	0.0/0.0	12.2/100.0	12.2/66.7	43.9/100.0
LongRanger	30.8/92.3	27.3/38.7	26.8/41.4	33.3/93.3
NAIBR	93.0/100.0	75.6/64.8	73.2/71.1	95.4/95.2
LEVIATHAN	0.0/0.0	25.3/33.4	19.0/35.1	54.6/54.6
LEVIATHAN (IG)	0.0/0.0	58.1/69.4	35.7/73.0	97.6/100.0

Table S17: Sensitivity/precision (%) of structural variant (SV) callers to detect insertions of the SV length category A (50 - 300 bp) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequencing coverage			
Manta	79.3/100	95.7/100.0	87.2/100.0	94.1/100.0
Novel-X	0.0/0.0	0.7/100.0	2.8/100.0	50.9/98.5
Novel-X (2 BND)[1]	0.0/0.0	0.7/83.3	2.3/79.3	50.3/97.3
	90x sequencing coverage			
Manta	89.89/100	98.8/100.0	94.6/100.0	94.5/100.0
Novel-X	7.21/100	39.4/98.4	35.3/98.6	76.0/99.1
Novel-X (2 BND)[1]	6.81/95.8	38.8/97.5	34.9/97.7	75.28/98.0
	135x sequencing coverage			
Manta	89.5/100	99.0/100.0	94.7/100.0	94.3/100.0
Novel-X	29.6/98.7	63.6/98.9	53.4/98.2	78.7/98.9
Novel-X (2 BND)[1]	29.5/97.0	62.7/97.5	53.1/97.4	78.1/97.5
	180x sequencing coverage			
Manta	89.1/100	98.8/100.0	94.4/100.0	94.2/100.0
Novel-X	47.6/99.4	72.3/98.1	63.9/98.0	79.0/98.6
Novel-X (2 BND)[1]	47.2/98.3	71.9/97.1	63.3/97.2	77.9/97.4

[1]Start and end break points were considered for evaluation

Table S18: Sensitivity/precision (%) of structural variant (SV) callers to detect insertions of the SV length category B (0.3 - 5 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequencing coverage			
Manta	74.5/100.0	96.0/100.0	83.8/100.0	86.3/100.0
Novel-X	4.2/100.0	14.3/99.3	14.7/99.3	96.8/99.0
Novel-X (2 BND)[1]	3.6/90.7	12.2/84.7	11.4/75.3	95.8/97.7
	90x sequencing coverage			
Manta	80.3/100.0	98.6/100.0	88.7/100.0	85.3/100.0
Novel-X	83.0/99.2	97.0/98.9	90.7/98.7	97.2/99.6
Novel-X (2 BND)[1]	81.6/97.6	95.8/97.1	89.2/97.1	95.9/98.3
	135x sequencing coverage			
Manta	76.5/100.0	98.5/100.0	88.3/100.0	83.9/100.0
Novel-X	95.6/99.3	97.6/99.2	96.5/99.3	97.7/99.5
Novel-X (2 BND)[1]	94.0/97.9	96.2/97.7	95.2/98.0	96.4/98.2
	180x sequencing coverage			
Manta	74.6/100.0	98.7/100.0	87.5/100.0	82.9/100.0
Novel-X	96.8/99.3	97.5/99.0	97.2/99.0	95.4/99.4
Novel-X (2 BND)[1]	95.3/97.7	96.5/97.5	95.9/97.7	93.8/97.9

[1]Start and end break points were considered for evaluation

Table S19: Sensitivity/precision (%) of structural variant (SV) callers to detect deletions of the SV length category C (5 - 50 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequence coverage			
Manta	72.9/100.0	93.8/100.0	83.0/100.0	86.5/100.0
Novel-X	9.0/98.9	11.6/99.1	12.8/98.4	96.1/99.2
Novel-X (2 BND)[1]	2.4/26.8	2.7/23.6	2.9/21.5	95.7/98.7
	90x sequence coverage			
Manta	80.7/100.0	97.7/100.0	89.0/100.0	86.0/100.0
Novel-X	93.1/98.6	97.5/98.0	94.9/97.3	96.2/99.6
Novel-X (2 BND)[1]	86.6/91.5	95.0/95.5	90.9/93.2	96.0/99.0
	135x sequence coverage			
Manta	76.7/100.0	97.7/100.0	88.2/100.0	83.7/100.0
Novel-X	96.2/98.9	97.1/98.5	96.7/98.7	96.4/99.3
Novel-X (2 BND)[1]	94.7/97.8	96.5/97.7	95.9/97.9	96.1/98.9
	180x sequence coverage			
Manta	73.3/100.0	98.1/100.0	87.2/100.0	82.2/100.0
Novel-X	96.2/99.1	96.9/98.9	96.7/98.8	96.2/99.3
Novel-X (2 BND)[1]	95.5/98.6	96.5/98.4	96.1/98.3	95.5/99.0

[1]Start and end break points were considered for evaluation

Table S20: Sensitivity/precision (%) of structural variant (SV) callers to detect insertions of the SV length category D (50 - 250 kb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequencing coverage			
Manta	73.7/100.0	84.9/76.5	74.8/79.6	88.9/100.0
Novel-X	11.1/100.0	30.3/85.4	23.2/79.3	96.0/99.0
Novel-X (2 BND)[1]	0.0/0.0	0.0/0.0	1.0/2.5	96.0/99.0
	90x sequencing coverage			
Manta	79.8/100.0	83.8/76.2	85.9/81.5	84.9/100.0
Novel-X	96.0/97.9	75.8/75.8	79.8/80.6	96.0/99.0
Novel-X (2 BND)[1]	44.4/46.3	59.6/59.6	47.5/47.5	96.0/97.9
	135x sequencing coverage			
Manta	75.8/100.0	80.8/76.0	77.8/78.7	82.8/100.0
Novel-X	96.0/99.0	80.8/80.8	87.9/87.0	95.0/98.9
Novel-X (2 BND)[1]	88.9/90.7	76.8/76.0	83.8/85.0	93.9/97.9
	180x sequencing coverage			
Manta	71.7/100.0	83.8/76.9	75.8/79.8	78.8/100.0
Novel-X	96.0/99.0	80.8/81.6	87.9/87.0	92.9/98.9
Novel-X (2 BND)[1]	96.0/99.0	80.8/81.6	87.9/87.0	91.9/97.9

[1]Start and end break points were considered for evaluation

Table S21: Sensitivity/precision (%) of structural variant (SV) callers to detect insertions of the SV length category E (0.25 - 1 Mb) for four haplotype incidences (1/4, 2/4, 3/4, 4/4) and four sequencing coverages (45x, 90x, 135x, 180x).

	Haplotype incidence			
SV caller	1/4	2/4	3/4	4/4
	45x sequencing coverage			
Manta	65.3/100.0	71.4/72.9	73.5/80.0	89.8/100.0
Novel-X	10.2/100.0	20.4/69.7	20.4/69.8	93.9/98.0
Novel-X (2 BND)[1]	0.0/0.0	0.0/0.0	0.0/0.0	24.5/24.4
	90x sequencing coverage			
Manta	79.6/100.0	85.7/75.0	81.6/79.3	81.6/100.0
Novel-X	95.9/97.9	79.6/75.0	83.7/80.4	98.0/98.0
Novel-X (2 BND)[1]	4.1/4.3	10.2/9.6	17.3/17.3	71.4/70.0
	135x sequencing coverage			
Manta	79.6/100.0	81.6/72.4	73.5/78.3	81.6/100.0
Novel-X	93.9/96.0	77.6/74.5	77.6/77.6	98.0/100.0
Novel-X (2 BND)[1]	67.4/66.7	59.2/55.6	63.3/66.0	91.8/91.8
	180x sequencing coverage			
Manta	67.4/100.0	83.7/74.1	77.6/78.4	77.6/100.0
Novel-X	95.9/97.8	77.6/75.0	81.6/80.0	89.8/97.9
Novel-X (2 BND)[1]	89.8/91.8	71.4/72.2	71.4/70.0	87.8/95.7

[1]Start and end break points were considered for evaluation

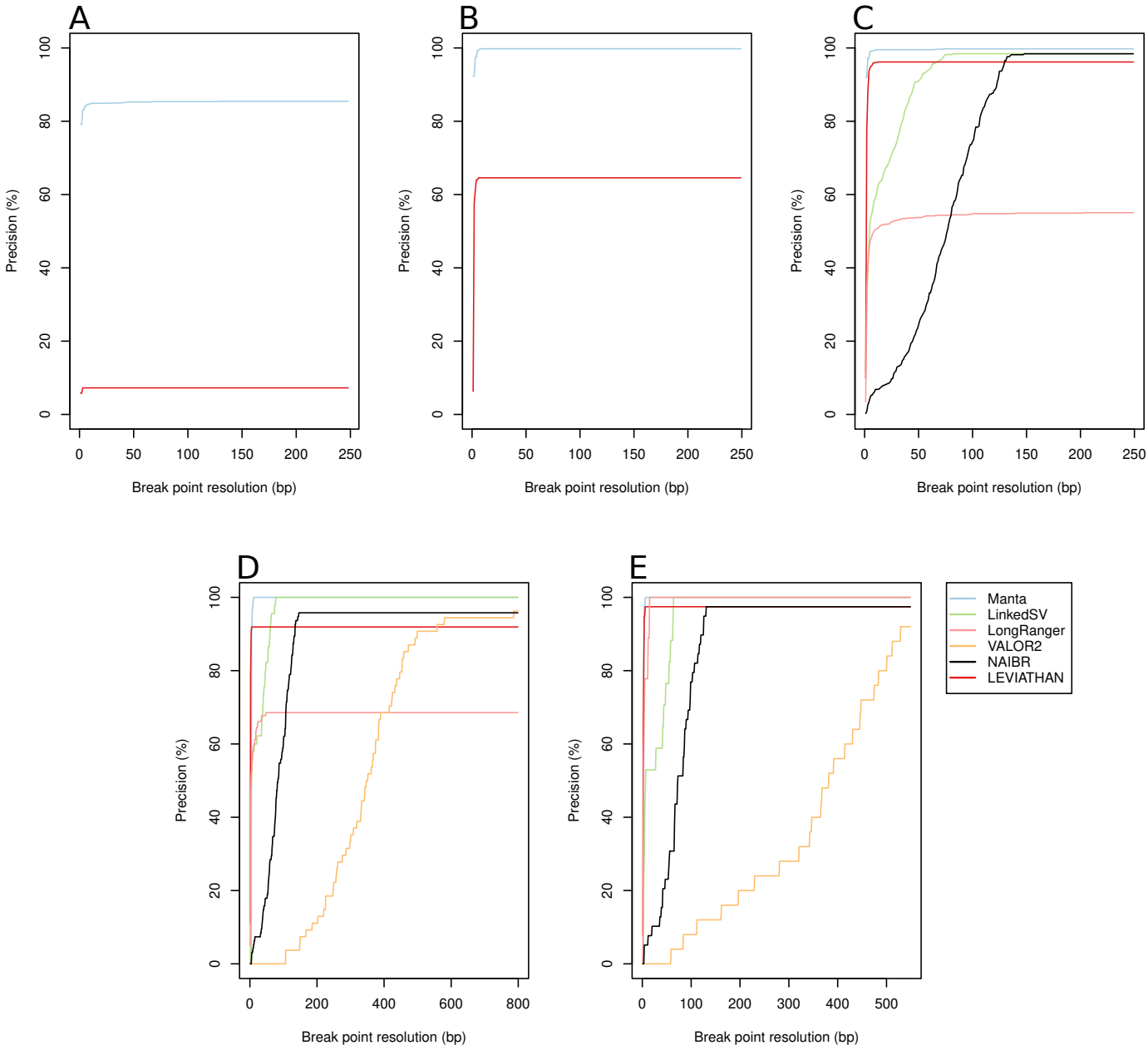

Fig. S1: Break point resolution in bp of the different SV callers for five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) based on the detection of homozygous (4/4) inversions using a linked-read sequencing coverage of 135x.

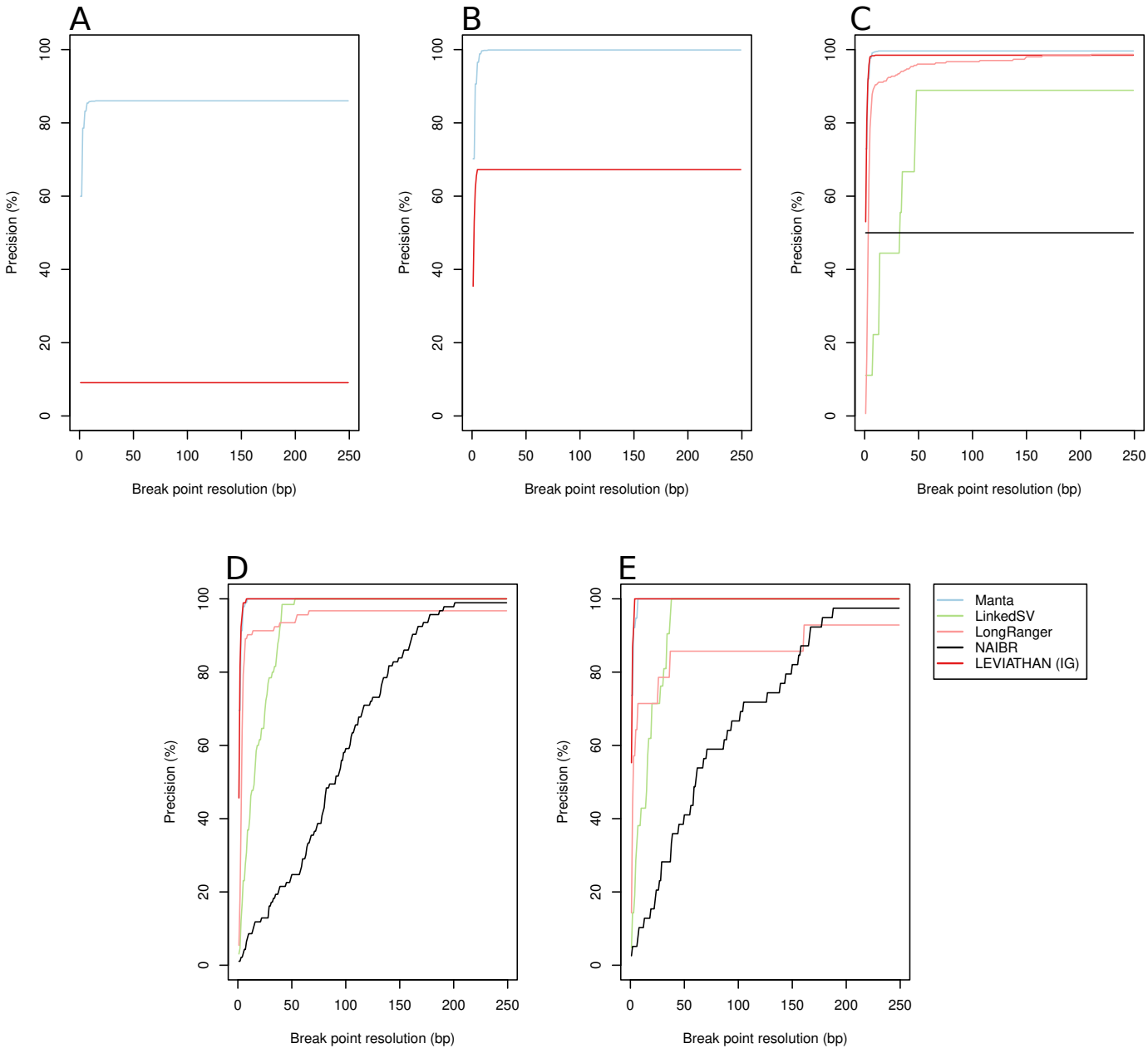

Fig. S2: Break point resolution in bp of the different SV callers for five structural variant (SV) length categories: A (50 - 300 bp), B (0.3 - 5 kb), C (5 - 50 kb), D (50 - 250 kb), E (0.25 - 1 Mb) based on the detection of homozygous (4/4) duplications using a linked-read sequencing coverage of 135x.

7. 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 submitted to PLOS Genetics in December, 2021 and is in peer review.

Authors:

Marius Weisweiler, Christopher Arlt, Po-Ya Wu, Delphine Van Inghelandt, Thomas Hartwig, Benjamin Stich

Contribution: First author

Marius Weisweiler and Benjamin Stich designated and coordinated the project.

Marius Weisweiler, Christopher Arlt, and Po-Ya Wu performed the data analyses.

Thomas Hartwig extracted DNA and prepared the libraries.

Delphine Van Inghelandt contributed phenotypic data.

Marius Weisweiler and Benjamin Stich wrote the manuscript.

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,**]

[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
*These authors contributed equally
**Corresponding author: benjamin.stich@hhu.de, Tel: **49-211/81-13395

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 (SNV). However, accurate while cost-efficient discovery of SV in complex genomes remains challenging. The objectives of our study were to (i) improve SV discovery 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 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 5kb 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 using data from single nucleotide polymorphism 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.

AUTHOR SUMMARY

Larger genomic rearrangements are defined as structural variants (SV) which cover at least 50bp, including deletions, insertions, inversions, duplications, and translocations. These SV are a leading cause for phenotypic variation. Due to the development of new sequencing technologies and corresponding bioinformatic algorithms, it is now possible to detect SV in complex cereal genomes in addition to examine single nucleotide variants. Therefore, we investigated based on computer simulations the sensitivity and precision to detect SV using short-read sequencing technology and used the best combination of SV callers to detect SV in the barley genome. Further, we could associate SV with gene expression and observed for the SV high abilities to predict important agronomic traits. Our findings have the potential to lead to increased efficiency of prediction-based breeding programs.

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 *Arabidopsis thaliana* as first plant species (The Arabidopsis Genome Initiative, 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 Pacbio or Nanopore 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 Aluminium 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). 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) improve SV discovery 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 resource for mapping quantitative traits, the double round robin population (Casale et al., 2021), (iii) quantify the association of SV clusters with transcript abundance, and (iv) evaluate the use of SV clusters for the prediction of quantitative phenotypic traits.

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 SV types (deletions, insertions, duplications, inversions, and translocations) in five SV length categories (A: 50 - 300bp; B: 0.3 - 5kb; C: 5 - 50kb; D: 50 - 250kb; E: 0.25 - 1Mb) 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 to 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 to 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 small insertions and deletions (2 - 49bp, INDELs) by 3% compared to using the single SV callers (Table 3). With 6%, an even higher difference for the sensitivity to detect translocations was observed between the combination of Manta | GRIDSS | (Delly & Lumpy) and single SV callers.

As a next step, 65x sequencing coverage was 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 (Fig. 1). 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. 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 to 7% higher based on Illumina sequencing data than based on linked-read sequencing data.

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

Across the 23 barley inbreds, we detected 629,974 SV clusters using the best combination of SV callers (Table 4). These comprised 313,061 deletions, 70,674 insertions, 96,541 duplications, 6,876 inversions, and 142,822 translocations. Additionally, 13,932,338 INDELs were detected across the seven chromosomes. The proportion of SV clusters which were annotated as transposable elements varied from 0.4% for inversions to 53.3% for translocations.

We performed a PCR based validation for detected deletions and insertions (Supplementary Table S1, Supplementary Fig. S1). Six out of six deletions and five out of five insertions up to 0.3kb could be validated (Supplementary Fig. S2). Additionally, we could validate 11 out of 14 deletions between 0.3kb and 460kb (Supplementary Fig. S3), 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 50,000 to more than 100,000 (Fig. 2A). We observed a significant ($P < 0.05$) positive correlation ($r = 0.43$) between the sequencing coverage, calculated based on 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 vs. variety status of the inbreds. In contrast, principal component analyses based on presence/absence matrices of the SV clusters revealed a clustering of inbreds by spike morphology, geographical origin, and landrace vs. variety status (Supplementary Fig. S4).

Out of the 629,974 SV clusters, 56% (353,278) appeared in only one of the 23 inbreds, whereas 17% (105,157) were detected in at least five inbreds (Fig. 2B, Supplementary Fig. S5). Additional analyses revealed a significant although weak correlation ($r = 0.0369$, $P = 1.16\text{x}10^{-136}$) between the length of a SV cluster and its minor allele frequency (MAF). The average MAF of SV clusters with a length of 250kb to 1Mb was 0.09, while that of SV clusters with a length of 50 - 250kb or 50bp - 50kb was 0.10 and 0.12, respectively (Fig. 3). SV clusters annotated as transposable elements had a shorter average length of 4,622bp and a higher MAF of 0.17 compared to SV clusters that were not annotated as transposable elements (7,310bp, 0.14). The average MAF of the individual SV types was the highest for insertions with 0.16, followed by deletions, inversions, duplications, and translocations with values of 0.14, 0.10, 0.10, and 0.09, 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 = 0.5375, P < 1x10^{-15}) of nucleotide diversity (π) of SV clusters and SNV, measured in 100kb windows along the seven chromosomes. 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. 4). For all centromeres, a significantly (P < 5.38x10^{-20}) 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 14.1% considerably lower compared to what is expected based on the physical length of these regions (25.6%). Only 1.4% of all detected SV hotspots were observed in pericentromeric regions. Compared to the five SV types, the genome-wide distribution of INDELs was more equal. Their occurrences peaked not only within, but also distal to pericentromeric and centromeric 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 linkage disequilibrium (LD) between each SV cluster and SNV located within 1kb and compared this with the extent of LD between the closest SNV to the SV cluster and the SNV within 1kb. Across the different SV types, 32.7-79.6% have at least one SNV within 1kb that showed an $r^2 \geq 0.6$ (Table 5). In contrast, 83.2-90.8% of SNV that are closest to the SV cluster showed an $r^2 \geq 0.6$ to another SNV within 1kb.

In the next step, we examined the presence of SV clusters relative to the position of genes. The highest proportion of SV clusters (63%) was located in intergenic regions of the genome (Fig. 5). The second largest fraction (27%) of SV clusters was present in the 5kb up- or downstream regions of genes, which is considerably higher compared to that of INDELs (15%) and SNV (7%). Within the group of "5kb up- or downstream to genes"- SV clusters, 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 15% lower than 5kb upstream of the TSS (Fig. 6). A similar trend was observed for the 5kb downstream regions ($\sim$ 10%).

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.

Selection on SV clusters

For landraces, as well as cultivars, significant ($P < 0.05$) leftward shifts of the unfolded site frequency spectrum (SFS) for all SV types compared to synonymous SNV were observed (Fig. 7A, 7B). A particularly high fraction (∼30-40%) of duplications and translocations showed a low derived allele frequency. To quantify the strength of selection that acts on the different types of variants, we estimated their fitness effects based on the SFS and with synonymous SNV used as neutral control. For more than 80% of insertions, duplications, inversions, and translocations a fitness effect of < -100 and, thus, a sign of selection, was observed in landraces as well as in cultivars. Interestingly, the proportion of deletions with fitness effects < -1 was higher for cultivars than for landraces.

Association of SV clusters with gene expression

We evaluated the strength of the association of the allele distribution at SV clusters with gene expression variation across 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 (Fig. 8, Supplementary Fig. S6). 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. S7). In addition, we investigated

a possible association between SV clusters and gene expression on an individual gene basis. For a total of 2,546 out of 30,156 gene-associated SV clusters a significant ($P < 0.05$) association with the gene expression of the associated gene was observed (Fig. 9). The mean Tajima's D for gene-associated SV clusters that were significantly associated with gene expression was with 1.111 slightly higher than those for which no significant association with gene expression was observed (1.099). Both values were considerably higher compared to those means of intergenic SV clusters (-0.258).

Prediction of phenotypic variation from SV clusters

The prediction ability of seven quantitative phenotypic traits using SV clusters as well as SNV from a single nucleotide polymorhpism (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 0.637. The SV clusters had the highest prediction power, followed by SNV&INDELs, SNP array, and gene expression in decreasing order (Fig. 10). Compared to these differences, those among the median prediction abilities of the different SV types were small. The highest ability was observed for insertions and the lowest for translocations. 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. S8). The mean of the optimal weight across the seven traits was highest for gene expression (0.41) and lowest for SV clusters (0.21) (Table 6).

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 exaggerated if many inbreds with large and complex genomes 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 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.

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

The costs for creating linked-read sequencing libraries are considerably higher compared to those 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 (Fig. 1). One reason might be that the SV callers used in our study do not fully exploit linked-read data. Indeed, 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 (Fig. 1). 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 more 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.

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

Validation of SV in the barley genome

In a first step, we explored whether known SV can be recovered in our data set. Taketa et al. (2008) discovered a 17kb 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 17kb deletion on chromosome 7H. In contrast, the deletion was absent in the 21 other parental inbreds.

In the next step, a PCR based approach was used to validate a 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. S1) for deletions and insertions up to 0.3kb (Supplementary Fig. S2). 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 eleven out of 14 deletions between 0.3kb and 460kb (Supplementary Fig. S3) 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. S3B). 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 three additional cases, we verified the absence of the predicted deletion sequence in the non-reference genotype but found no evidence of the presence of the same sequence in Morex (Supplementary Fig. S3A). This could be explained by an error in that version of the reference sequence v2 (Monat et al., 2019) that was used for our analyses. BLAST results for these three SV using the recently released Morex reference sequence v3 (Mascher et al., 2021) support our PCR results and suggest that by chance we had selected SV where the reference sequence v3 had corrected an error of v2.

In conclusion, for 22 of the 25 tested SV (Supplementary Table S1), 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 629,974 SV clusters (Table 4). This corresponds to 1 SV cluster every 6,769 bp and corresponds to what was observed by Jayakodi et al. (2020). This number is considerably higher than 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). The lower number of SV clusters detected in cucumber and peach might be explained by the usage of a lower number of SV callers (cf Zhang et al., 2015) and the focus on heterozygous SV clusters in the case of the peach study (cf 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 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 4). 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). This explanation is supported by the observation of a considerably higher frequency of duplications in repeat regions (8,929 bp^{-1}) compared to the rest of the genome (54,322 bp^{-1}).

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 14.1% of the SV clusters were located in pericentromeric regions, which make up 25.6% of the genome, whereas the rest was located distal of the pericentromeric regions (Fig. 4). This pattern was even more pronounced for SV hotspots, i.e. regions with a significant higher amount of SV clusters than expected by the average genome-wide distribution.

Almost all SV hotspots (98.4%) were located distal of the pericentromeric regions (74.4% of the genome) where higher recombination rates are observed. Our observation indicates that the majority of SV clusters in barley is caused by mutational mechanisms related to DNA recombination-, replication-, and/or repair-associated processes which has been discussed in the human genetic context (Carvalho and Lupski, 2016) and is only to a low extent due to the activity of transposable elements. This is supported by the observation that, with the exception of translocations, only 0.4 to 25.7% of SV clusters were located in genome regions annotated as transposable elements (Table 4). Similar results were observed in rice (Fuentes et al., 2019).

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

Association of SV clusters with transcript abundance

About 63% of the SV clusters were detected in the intergenic space (Fig. 5). The remaining SV clusters were gene-associated and detected in regions either 5kb up- or downstream of genes (∼27%) while ∼10% were detected in introns and exons (Fig. 5). 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.

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 (Fig. 8). In addition, 1,448 gene-associated SV clusters were individually associated ($P < 0.05$) with genome-wide gene expression. A further 2,546 alleles of gene-associated SV clusters were significantly ($P < 0.05$) associated with the expression of the corresponding 2,097 genes (Fig. 9). Additional support is given

by the observation that despite that SV clusters have a similar distribution across the genome as SNV, SV clusters covered more positions (in bp) of promoter regions than SNV (Fig. 6). 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. Gene-associated SV clusters had a higher average Tajima's D (1.111) compared to intergenic SV clusters (-0.258). This indicated the importance of balancing selection for the former SV clusters. We examined in detail the genes for which SV clusters were significantly associated with gene expression and a particularly high Tajima's D value was observed. The genes involved in resistance against or expressed in response to infection of powdery mildrew (Zimmermann et al., 2006; Li et al., 2013; Koch et al., 2017; Kumar et al., 2018; Galli et al., 2021; Velásquez-Zapata et al., 2021) were over-represented among them. The reason for this finding remains elusive.

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, 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 (Casale et al., 2021). The detailed characterization of the SV pattern of the parental inbreds, presented in this study, will therefore be an important information for the ongoing identification of candidate genes that underlay quantitative trait loci.

However, instead of examining the association of individual SV clusters with phenotypic traits, we evaluated their potential to predict seven phenotypic traits in comparison to 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. 10). 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 (Table 6). 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 (Table 5). 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 e.g. generated in our study for barley. However, this requires further research.

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). 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 - 300bp; B: 0.3 - 5kb; C: 5 - 50kb; D: 50 - 250kb; E: 0.25 - 1Mb) plus INDELs (2-49bp). Translocations were simulated for 50bp - 1Mb (ABCDE). We simulated SV with a mutation rate of $1.9\text{x}10^{-6}$ for the SV length categories A-C and INDELs, whereas mutation rates of $3.8\text{x}10^{-6}$ and $1.9\text{x}10^{-7}$ were assumed for SV length categories D and E, respectively. For each type of SV, we used BBMap's randomreads.sh (BBMap - Bushnell B. - `sourceforge.net/projects/bbmap/`) to simulate 2x150bp Illumina reads with a sequencing coverage of 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.

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) was 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. We calculated the sensitivity (1), precision (2), and the F1-score (3) as

$$\text{Sensitvity} = \text{TP}/(\text{TP}+\text{FN}) \tag{1}$$

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

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

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 $\leq$ 2bp from those of the simulated INDEL and the length did differ by $\leq$ 5bp. For SV length category A, a TP SV had break points that did differ $\leq$ 10bp from those of the simulated SV and the SV length did differ by $\leq$ 20bp. For the other SV length categories, a TP SV had break points and length differences compared to the simulated SV of $\leq$ 50bp. For insertions where no SV length was detected, the start of a TP insertion had a break point that did differ $\leq$ 10bp from this of the simulated insertion. For translocations, a TP translocation had break points that did differ $\leq$ 50bp 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 $\geq$ 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 $\leq$ 95% in at least one SV length category, SV calls were combined with a logical and ("&"). If the precision of the combinations of the second step increased to $\geq$ 95% in all SV length categories, SV calls of this combinations were kept for the particular SV type and were combined with a logical or with those of the first step.

The threshold of $\geq$ 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 S2) using the MSTRAT algorithm (Gouesnard, 2001). These inbreds are the parents of the double round robin population (Casale et al. 2021). Paired-end sequencing libraries with an insert size of 425bp were sequenced 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 v2 (Monat et al., 2019) 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, 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 allele (0/1) calls 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. 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 shortest SV was considered. 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 50bp and the difference of the two break points had to be smaller than 50bp 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 were called using GATK. First, GATK's HaplotypeCaller was used in single sample GVCF mode, afterwards 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 (QD $<$ 2.0; QUAL $<$ 30.0; SOR $>$ 3.0; FS $>$ 60.0; MQ $<$ 40.0; MQRankSum $<$ -12.5; ReadPosRankSum $<$ -8.0).

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: 5kb upstream/downstream gene-associated SV clusters were located in the 5kb 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 v2 (Monat et al., 2019). Deletions, duplications, inversions, INDELs, and insertions with known length were annotated as transposable elements if the reciprocal overlap was $\geq 80\%$ (Fuentes et al., 2019). Insertions with unknown length were classified as transposable elements if the detected break point of the insertion was inside the transpos-

able 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 1Mb 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).

SNV annotation: SIFT4G (version 2.4) was used to annotate and predict synonymous, non-synonymous (score > 0.05), and deleterious (score ≤ 0.05) SNV based on the conversion of amino acid sequences (Vaser et al., 2016). The SIFT4G database was built using SIFT4_Create_Genomic_DB with the uniref90 database, the Morex reference sequence v2, and its corresponding predicted genes and proteins.

Population genetic analyses: LD measured as r^2 (Hill and Robertson, 1968) was calculated between each SV type and linked SNV. Nucleotide diversity (π) was calculated in 100kb windows along the seven chromosomes separately for SV clusters and SNV using bcftools (version 1.10.2) (Danecek et al., 2021).

SFS of synonymous, non-synonymous, deleterious SNV, and all SV types was calculated for cultivars and landraces using three *Hordeum vulgare* subsp. *spontaneum* accessions (Li et al., 2020) as outgroup. For landraces and cultivars, the population size was reduced to ten, because of computational reasons, where those genotypes with the highest sequencing coverage have been selected. The distribution of fitness effects (DFE) was determined by polyDFE (version 2.0) (Tataru et al., 2017) using 500 iterations and the model A, based on the SFS for the SV types, non-synonymous, and deleterious SNV while those of the synonymous SNV were considered as neutral control. The results were presented with 95% confidence intervals obtained from 100 bootstrap runs. Tajima's D was calculated using vcftools (version 0.1.13) (Danecek et al., 2011) for each SV cluster as well as various subsets.

SV clusters and gene expression: SV clusters which were assigned into one of the gene-associated SV categories, namely 5kb 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 frag-

ments 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 MAF > 0.15, Pearson's correlation coefficient with the first three principal components was estimated, where presence and absence of SV clusters were used as metric character. 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 agro-ecologically diverse 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:

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

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} envi-

ronment, 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) 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. $\mathbf{W}$ was defined as a matrix of feature measurement for the respective predictor. The dimensions of $\mathbf{W}$ were the number of barley inbreds ($n = 23$) times the number of features in the corresponding predictor (m) ($m_{SNP\ array} = 38,025$, $m_{gene\ expression} = 67,844$, $m_{SNV\&INDELs} = 3,110,041$, $m_{SNV} = 2,373,586$, $m_{INDELs} = 736,455$, $m_{SVclusters} = 629,535$, $m_{deletions} = 312,704$, $m_{duplications} = 96,521$, $m_{insertions} = 70,618$, $m_{inversions} = 6,874$, $m_{translocations} = 142,818$). The additive relationship matrix $\mathbf{G}$ was defined as $\mathbf{G} = \frac{\mathbf{W}^*\mathbf{W}^{*\mathbf{T}}}{m}$, where $\mathbf{W}^*$ was a matrix of feature measurement for the respective predictor, whose columns are centered and standardized to unit variance of $\mathbf{W}$, and $\mathbf{W}^{*\mathbf{T}}$ was the transpose of $\mathbf{W}^*$.

Furthermore, to investigate the performance of a joined weighted relationship matrix (Schrag et al., 2018) to predict phenotypic variation, the three $\mathbf{G}$ 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.

Five-fold 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.

DECLARATIONS

Availability of data and materials

Raw DNA sequencing data of the 23 barley inbreds have been deposited into the NCBI Sequence Read Archive (SRA) under the accession PRJNA77700 and will become available after manuscript acceptance (`https://dataview.ncbi.nlm.nih.gov/object/PRJNA777004?reviewer=el83fbl241mgqmbjdireuafcic`). Raw mRNA sequencing data are available under the accession PRJNA534414. Raw DNA sequencing data of the three *Hordeum vulgare* subsp. *spontaneum* accessions are available under the accession PRJNA622206 (SRS6405919, SRS6405923, SRS6405924). Data of gene expression, SNP array, and adjusted entry means of phenotypes are available via figshare (`https://figshare.com/s/3eaf2d54df0c2496b68c`). Snakemake workflows are available via github (`https://github.com/mw-qggp/SV_barley`). Further scripts are available from the authors upon request.

Acknowledgements

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

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686 111). 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.

Authors' contributions

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.

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.

Competing interests

The authors declare that they have no competing interests.

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. Nature Reviews Genetics 12:363–376

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, 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, 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. Nature Methods 9:133–137

Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, Ramsay L, et al. (2017), Development and evaluation of a barley 50k iSelect SNP array. Frontiers in Plant Science 8:1792

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, et al. (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 Bioinformatics 10:421

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

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 Research 27:1–11

Carvalho CM, Lupski JR (2016), Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics 17:224–238

Casale F, Van Inghelandt D, Weisweiler M, Li J, Stich B (2021), Genomic prediction of the recombination rate variation in barley - a route to highly recombinogenic genotypes. Plant Biotechnology Journal https://doi:10.1111/pbi.13746

Chaisson MJ, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, et al. (2019), Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications 10:1784

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, et al. (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, et al. (2017), The impact of structural variation on human gene expression. Nature Genetics 49:692–699

Craig-Holmes AP, Moore FB, Shaw MW (1973), Polymorphism of human C-band heterochromatin. I. Frequency of variants. American Journal of Human Genetics 25:181–192

Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. (2001), The sequence of the human genome. Science 291:1304–1351

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, et al. (2011), The variant call format and VCFtools. Bioinformatics 27:2156–2158

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, et al. (2021), Twelve years of SAMtools and BCFtools. GigaScience 10:1–4

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

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

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, 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, et al. (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, et al. (2019), Structural variants in 3000 rice genomes. Genome Research 29:870–880

Galli M, Martiny E, Imani J, Kumar N, Koch A, Steinbrenner J, Kogel KH (2021), CRISPR/SpCas9-mediated double knockout of barley Microrchidia MORC1 and MORC6a reveals their strong involvement in plant immunity, transcriptional gene silencing and plant growth. Plant Biotechnology Journal https://doi.org/10.1111/pbi.13697

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, et al. (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. Briefings in bioinformatics 22:1–15

Gouesnard B (2001), MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. Journal of Heredity 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 Breeding 129:271–279

Hill WG, Robertson A (1968), Linkage disequilibrium among neutral genes in finite populations. Theoretical and Applied Genetics 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, 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 Biology 21:72

Koch A, Kang HG, Steinbrenner J, Dempsey DA, Klessig DF, Kogel KH (2017), MORC proteins: novel players in plant and animal health. Frontiers in Plant Science 8:1720

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. (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 Biology 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. Molecular Biology and Evolution 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

Kumar N, Galli M, Ordon J, Stuttmann J, Kogel KH, Imani J (2018), Further analysis of barley MORC1 using a highly efficient RNA-guided Cas9 gene-editing system. Plant Biotechnology Journal 16:1892–1903

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

Li DQ, Nair SS, Kumar R (2013), The MORC family: new epigenetic regulators of transcription and DNA damage response. Epigenetics 8:685–693

Li K, Ren X, Song X, Li X, Zhou Y, Harlev E, Sun D, Nevo E (2020), Incipient sympatric speciation in wild barley caused by geological-edaphic divergence. Life Science Alliance 3:e202000827

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

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, et al. (2020a), 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 (2020b), 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. Computational and Structural Biotechnology Journal 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 Biology 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, et al. (2019), Resolving the full spectrum of human genome variation using linked-reads. Genome Research 29:635–645

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, et al. (2013), Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proceedings of the National Academy of Sciences of the United States of America 110:5241–5246

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk

V, et al. (2017), A chromosome conformation capture ordered sequence of the barley genome. Nature Publishing Group 544:427–433

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, et al. (2021), Long-read sequence assembly: a technical evaluation in barley. The Plant Cell 33:1888–1906

McColgan P, Tabrizi SJ (2018), Huntington's disease: a clinical review. European Journal of Neurology 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. International Journal of Cancer 24:27–33

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, et al. (2019), TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biology 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, et al. (2013), Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biology 14:R58

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, et al. (2014), Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 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. Molecular Breeding 31:27–37

Nowell P, Hungerford D (1960), Chromosome studies on normal and leukemic human leukocytes. Journal of the National Cancer Institute 25:85–109

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, et al. (2017), Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 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. Proceedings of the National Academy of Sciences 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, et al. (2017), Parkinson's disease associated with pure ATXN10 repeat expansion. npj Parkinson's Disease 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, et al. (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, et al. (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, et al. (2008), Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. Proceedings of the National Academy of Sciences of the United States of America 105:4062–4067

Tataru P, Mollion M, Glémin S, Bataillon T (2017), Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119

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 Research 40:e115

VanRaden P (2008), Efficient methods to compute genomic predictions. Journal of Dairy Science 91:4414–4423

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016), SIFT missense predictions for genomes. Nature Protocols 11:1–9

Velásquez-Zapata V, Elmore JM, Banerjee S, Dorman KS, Wise RP (2021), Next-generation yeast-two-hybrid analysis with Y2H-SCORES identifies novel interactors of the MLA immune receptor. PLOS Computational Biology 17:e1008890

Wang O, Chin R, Cheng X, Yan Wu MK, Mao Q, Tang J, Sun Y, et al. (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 Research 29:798–808

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, et al. (2018), Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017), Direct determination of diploid genome sequences. Genome Research 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, et al. (2012), Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology 30:105–111

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, et al. (2019), Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nature Genetics 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, et al. (2015), Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. The 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. Nature Biotechnology 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. Nature Plants 5:965–979

Zimmermann G, Bäumlein H, Mock HP, Himmelbach A, Schweizer P (2006), The multigene family encoding germin-like proteins of barley. Regulation and function in basal host resistance. Plant Physiology 142:181–192

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 - 300bp), B (0.3 - 5kb), C (5 - 50kb), D (50 - 250kb), and E (0.25 - 1Mb).

	SV length category				
SV caller	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

Table 3: Sensitivity/precision (%) of structural variant (SV) callers and combinations of them (for details see Material & Methods) to identify small insertions and deletions (2 - 49bp, INDELs) and translocations (50bp - 1Mb).

SV caller	Deletions (2 - 49bp)	Insertions (2 - 49bp)	Translocations (50bp - 1Mb)
Delly			85.6/76.0
Manta			89.4/100.0
Lumpy			83.2/82.4
GRIDSS	68.0/99.3	64.6/98.9	87.2/100.0
Pindel	92.4/97.9	87.5/98.7	
GATK	92.3/97.6	94.6/98.7	
Combination	95.5/98.9	94.8/98.7	95.4/99.8

Table 4: Summary of detected structural variants (SV) and small insertions and deletions (2 - 49bp, 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 v2.

SV type	Number of SV calls	Number of SV clusters		
			MAF > 0.05	TE
Deletions	1,042,134	313,061	166,920	10,691
Insertions	242,915	70,674	40,945	287 (18,173)[1]
Duplications	212,435	96,541	37,292	5,705
Inversions	15,955	6,876	2,572	30
Translocations	297,660	142,822	52,986	0 (74,621)[1]
INDELs	63,367,764	13,932,338	8,904,504	35

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

Table 5: Percentage of structural variant (SV) clusters or their closest neighboring single nucleotide variant (SNV) that show a maximum linkage disequilibrium (LD) estimate r^2_{max} to all SNV 1kb up and downstream of it. LD was calculated for three categories of minor allele frequencies (MAF) for SV clusters and the corresponding closest SNV.

		MAF					
Proportion [%] of r^2_{max}	r^2	[0,0.2)	[0.2,0.4)	[0.4,0.5)	[0,0.2)	[0.2,0.4)	[0.4,0.5)
		Between SV cluster and SNV			Between closest SNV to SV cluster and SNV		
	[1.0,0.8]	0.00	0.89	60.96	0.00	2.3	76.38
	(0.8,0.6]	47.18	70.74	13.34	89.50	87.80	12.84
Deletions	(0.6,0.4]	6.71	11.79	12.30	10.50	9.90	10.73
	(0.4,0.2]	36.52	8.21	7.54	0.00	0.00	0.00
	(0.2,0]	9.59	8.37	5.87	0.00	0.00	0.00
	[1.0,0.8]	0.00	0.89	59.57	0.00	2.16	76.76
	(0.8,0.6]	42.97	66.85	13.16	90.41	88.51	12.66
Insertions	(0.6,0.4]	8.02	12.83	12.57	9.59	9.33	10.51
	(0.4,0.2]	38.59	9.25	8.05	0.00	0.00	0.00
	(0.2,0]	10.42	10.18	6.65	0.00	0.00	0.00
	[1.0,0.8]	0.00	1.72	52.17	0.00	4.62	65.21
	(0.8,0.6]	32.68	60.81	14.13	83.32	78.84	16.31
Duplications	(0.6,0.4]	5.00	14.64	15.18	16.68	16.54	18.24
	(0.4,0.2]	50.08	11.24	10.54	0.00	0.00	0.00
	(0.2,0]	12.25	11.59	7.97	0.00	0.00	0.00
	[1.0,0.8]	0.00	1.47	48.32	0.00	3.18	66.14
	(0.8,0.6]	34.05	61.29	15.18	85.21	80.77	14.54
Inversions	(0.6,0.4]	6.15	15.96	15.89	14.79	16.05	19.19
	(0.4,0.2]	50.72	11.51	12.08	0.00	0.00	0.00
	(0.2,0]	9.08	9.78	8.52	0.00	0.00	0.00

Table 6: The optimal weights of the three predictors single nucleotide variants (SNV) and small insertions and deletions (2 - 49bp, INDELs, SNV&INDELs), structural variant (SV) clusters and gene expression that resulted in the highest prediction abilities for the seven traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), and thousand grain weight (TGW).

Traits	SNV&INDELs	SV clusters	Gene expression	Prediction ability
HT	0.0	0.1	0.9	0.63
LA	0.0	0.4	0.8	0.79
PH	0.1	0.0	0.9	0.54
SA	0.9	0.0	0.1	0.74
SL	0.5	0.1	0.4	0.70
SW	0.1	0.9	0.0	0.75
TGW	1.0	0.0	0.0	0.86
Mean (median)	0.37 (0.1)	0.21 (0.1)	0.41 (0.4)	

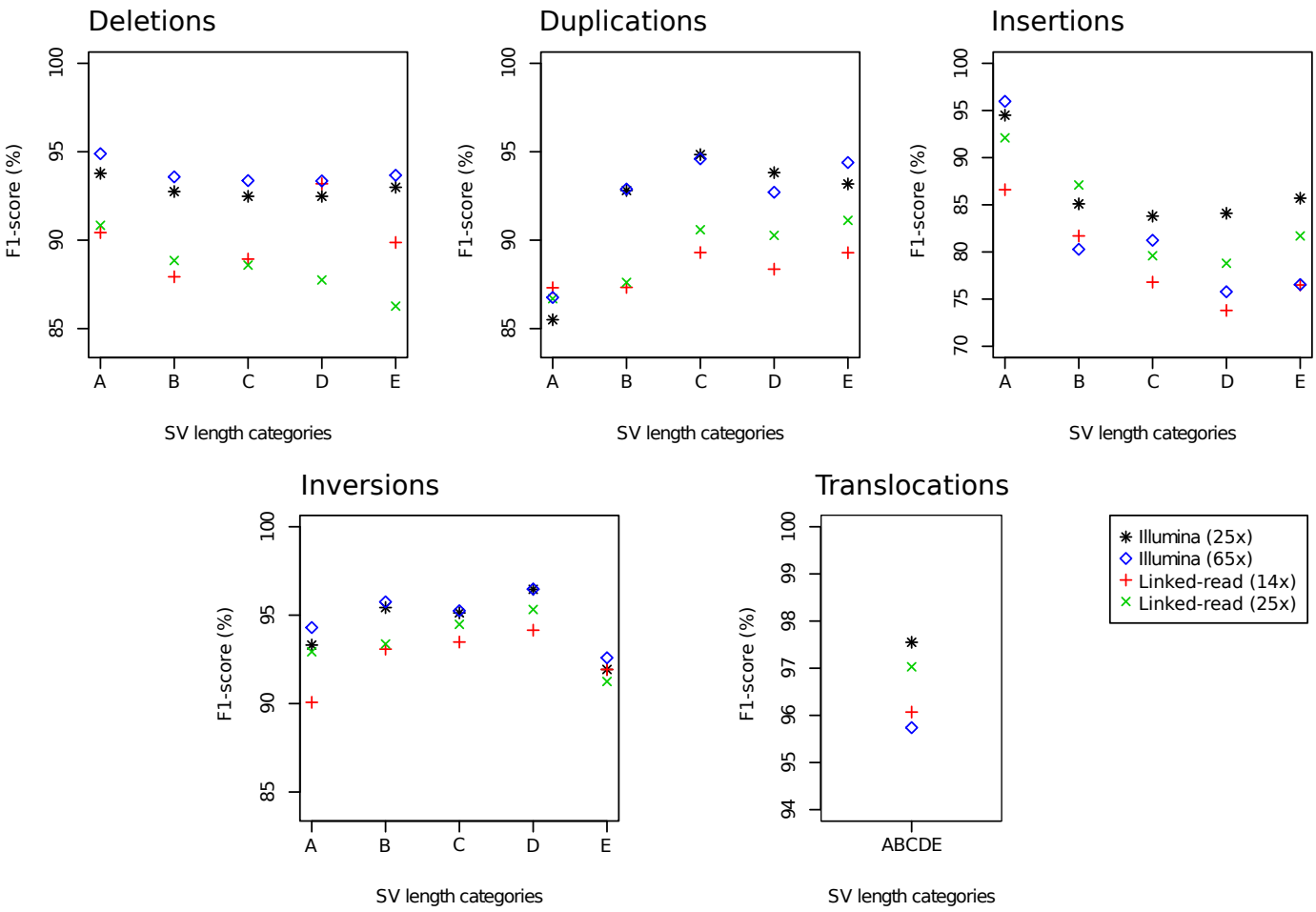

Fig. 1: F1-score, which is the harmonic mean of the precision and sensitivity, for the detection of deletions, duplications, insertions, inversions, and translocations of five structural variant (SV) length categories: A (50 - 300bp), B (0.3 - 5kb), C (5 - 50kb), D (50 - 250kb), E (0.25 - 1Mb) using the best combination of SV callers (for details see Material & Methods) based on 25x and 65x Illumina short-read sequencing as well as based on 14x and 25x linked-read sequencing coverage.

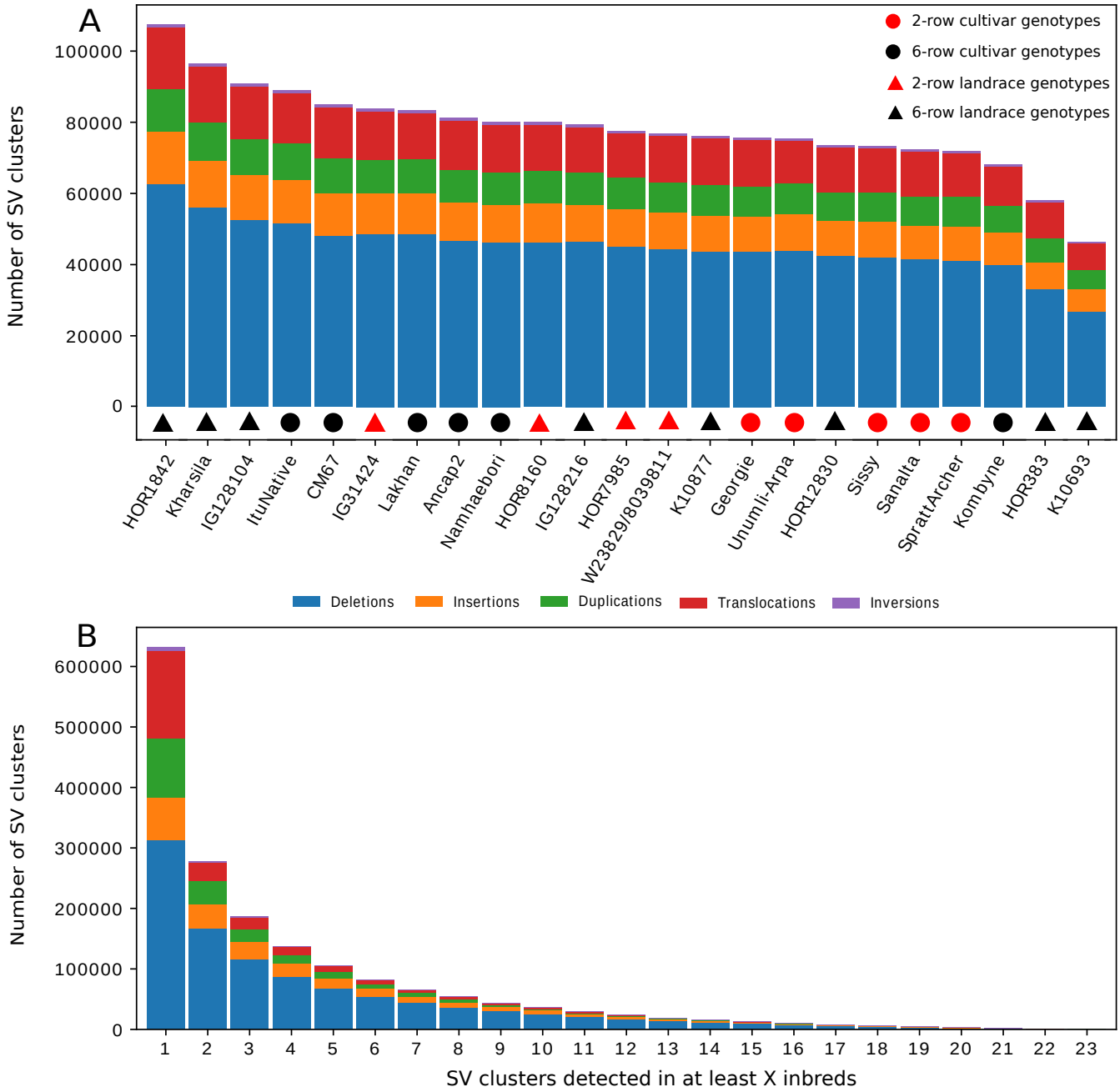

Fig. 2: Stacked bar graph of the number of different types of structural variant (SV) clusters detected in the 23 inbreds (A) and SV clusters which were detected in at least the given number of the inbreds (B).

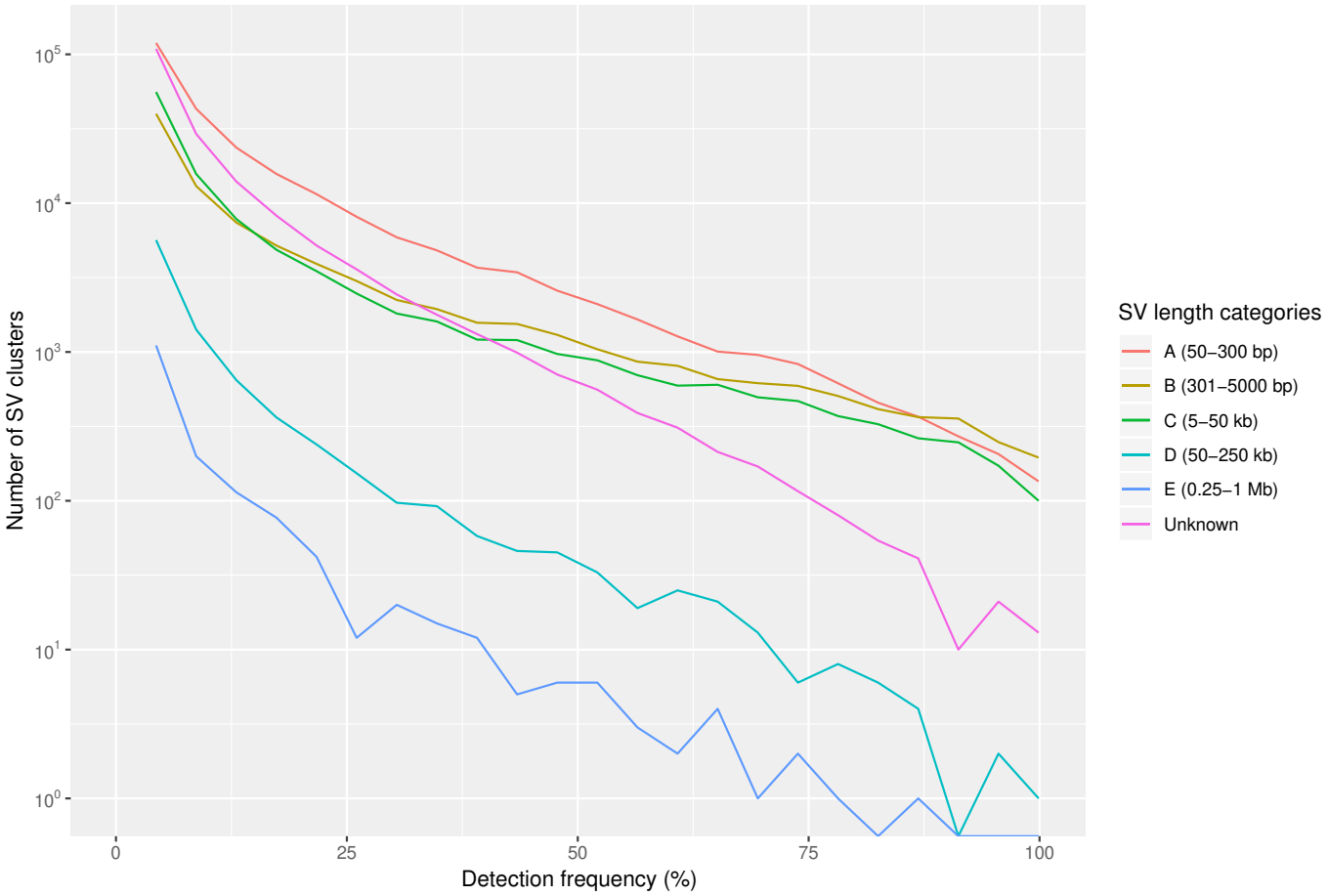

Fig. 3: Detection frequencies of structural variant (SV) clusters of different length categories across the 23 barley inbreds.

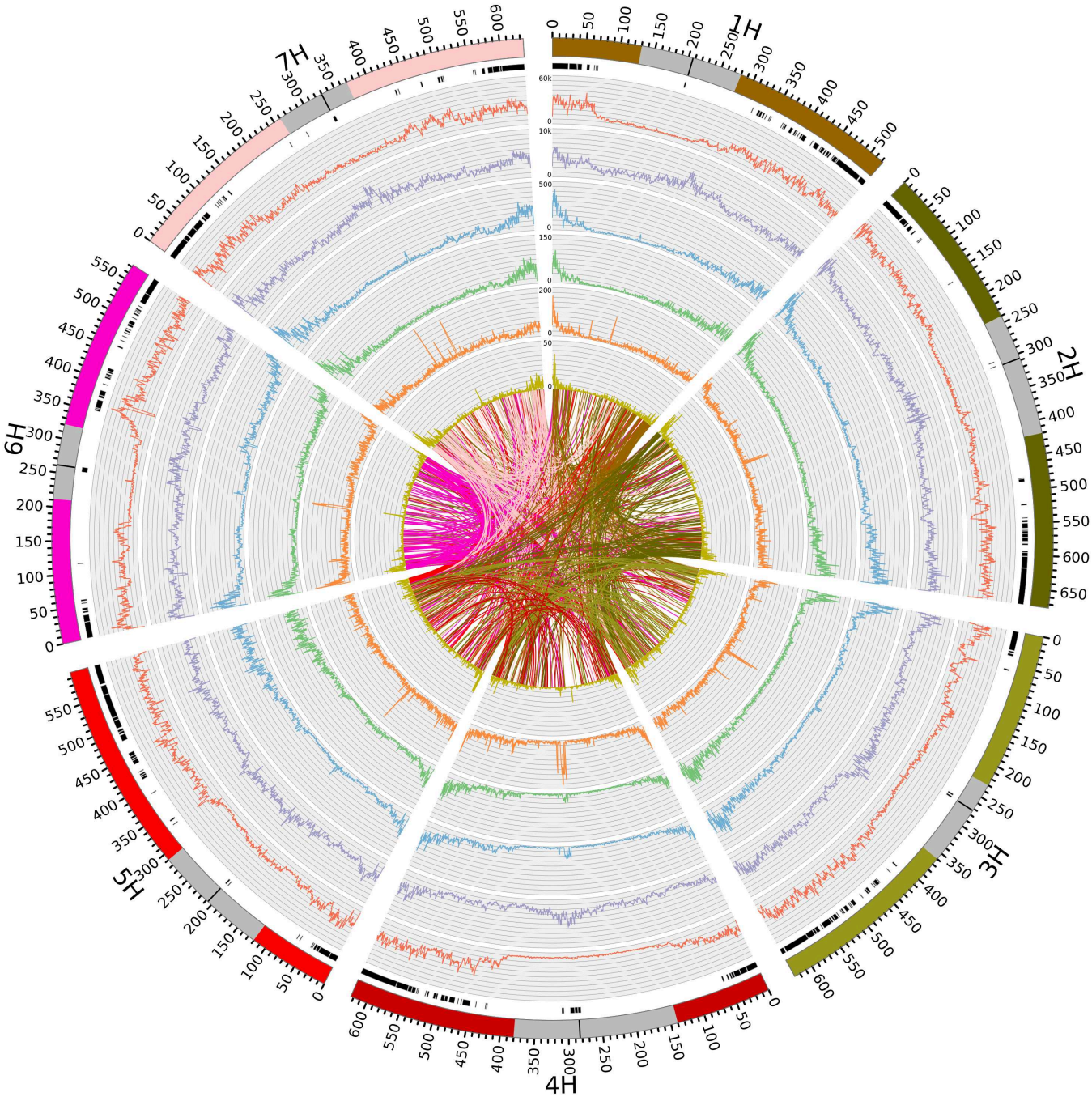

Fig. 4: Distribution of genomic variants among 23 barley inbreds across the seven chromosomes. The outermost circle denotes the chromosome number, the physical position, and as gray bar the peri-centromeric regions (Casale et al. 2021) plus the centromeres (black) according to the Morex reference sequence v2. The next inner circles report the SV cluster hotspots (black bars), frequencies of single nucleotide variants (red), small insertions and deletions (2 - 49bp, INDELs, purple), deletions (blue), insertions (green), duplications (orange), and inversions (yellow) which were detected among the 23 inbreds. Interchromosomal translocations that were observed in at least twelve inbreds are represented in the middle of the circle as colored lines connecting the two genomic regions.

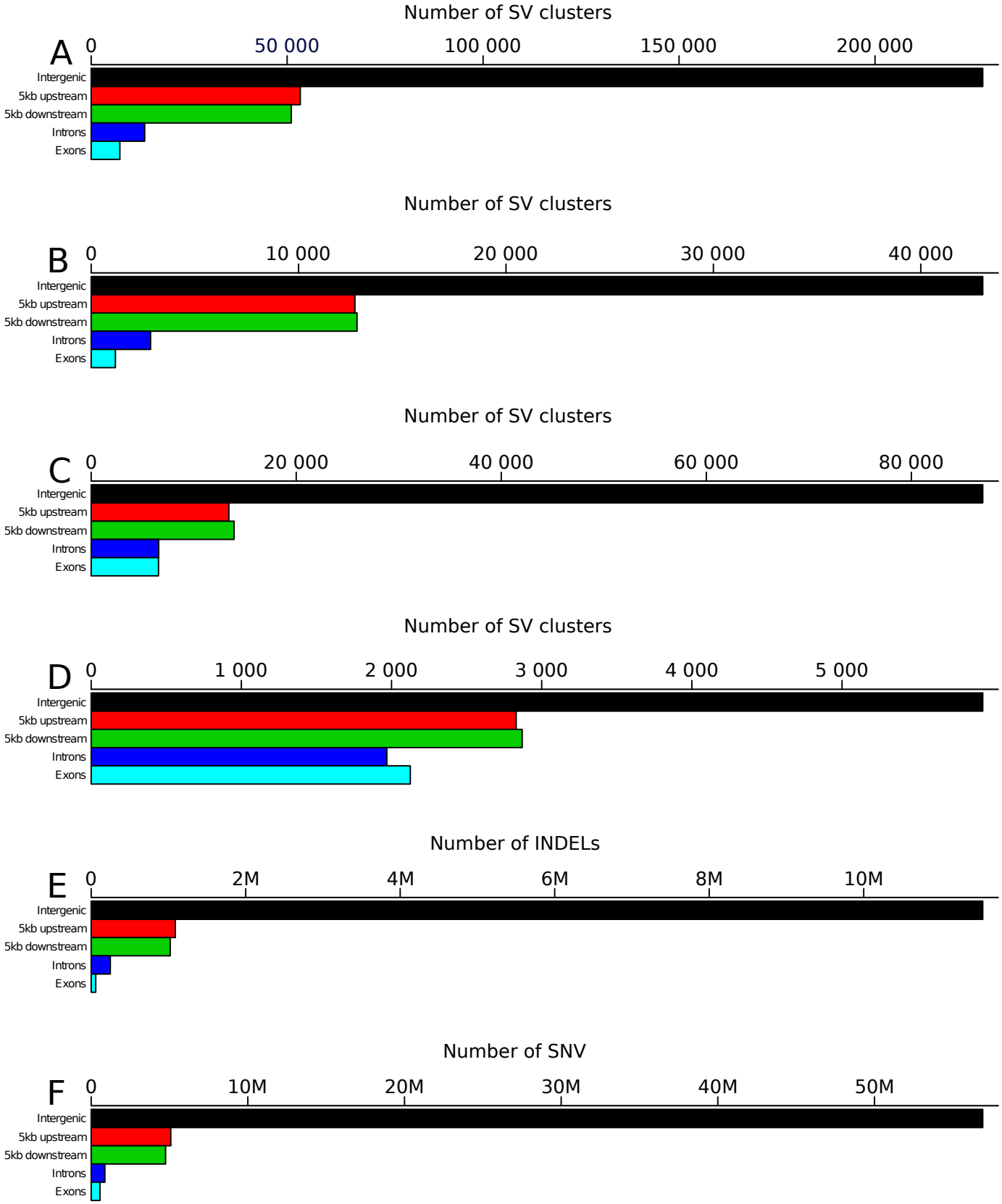

Fig. 5: The occurrence of deletions (A), insertions (B), duplications (C), inversions (D), small insertions and deletions (2 - 49bp, INDELs, E), and single nucleotide variants (SNV) (F) in five genomic regions.

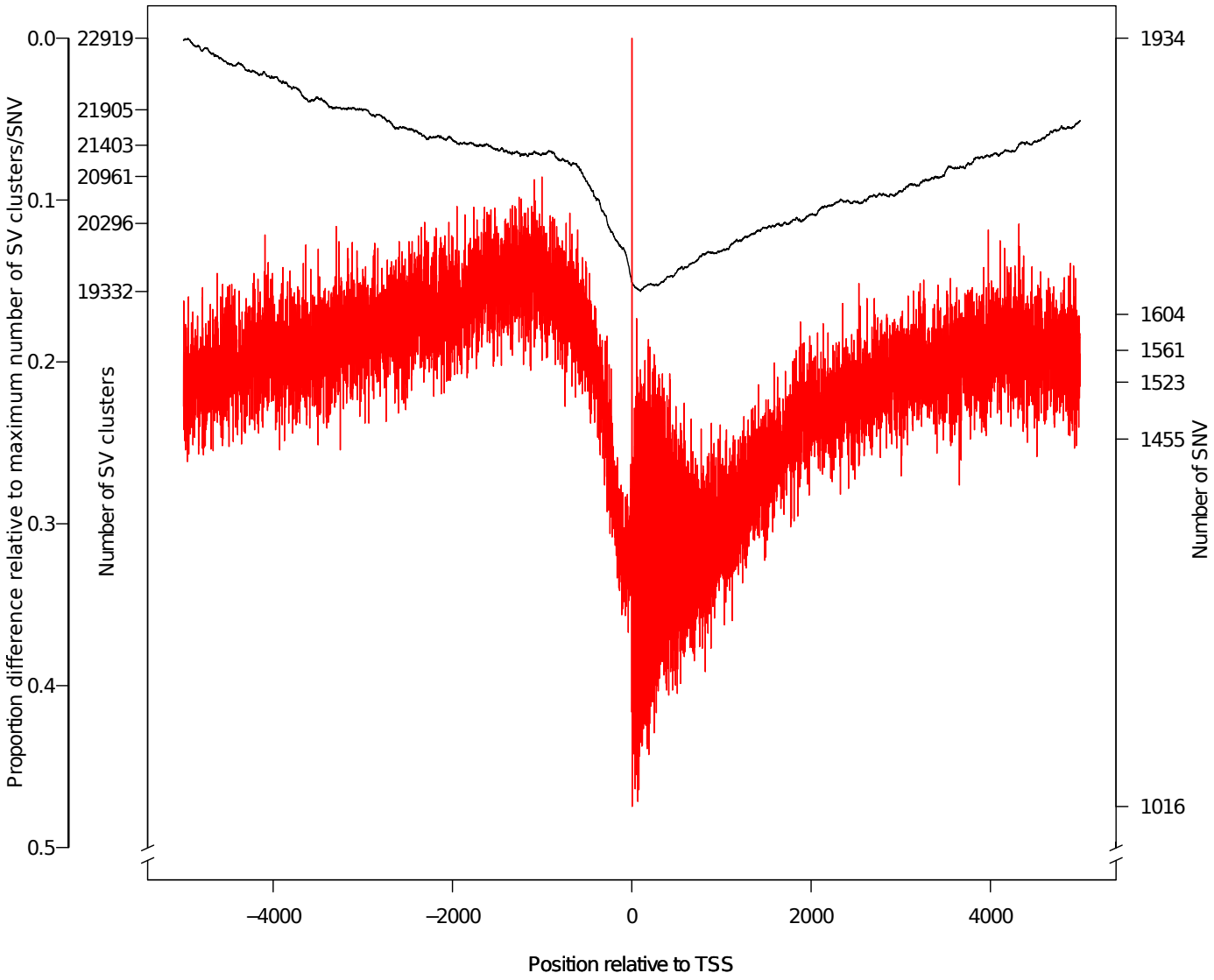

Fig. 6: Distribution of structural variant (SV) clusters (black) and single nucleotide variants (SNV, red) among 23 barley inbreds relative to the transcription start site (TSS) of a gene (x-axis). SV clusters and SNV were counted for every position from 5kb up- and downstream around the TSS of all genes (y-axes). As third y-axis, the proportion difference relative to the maximum number of SV clusters/SNV is illustrated.

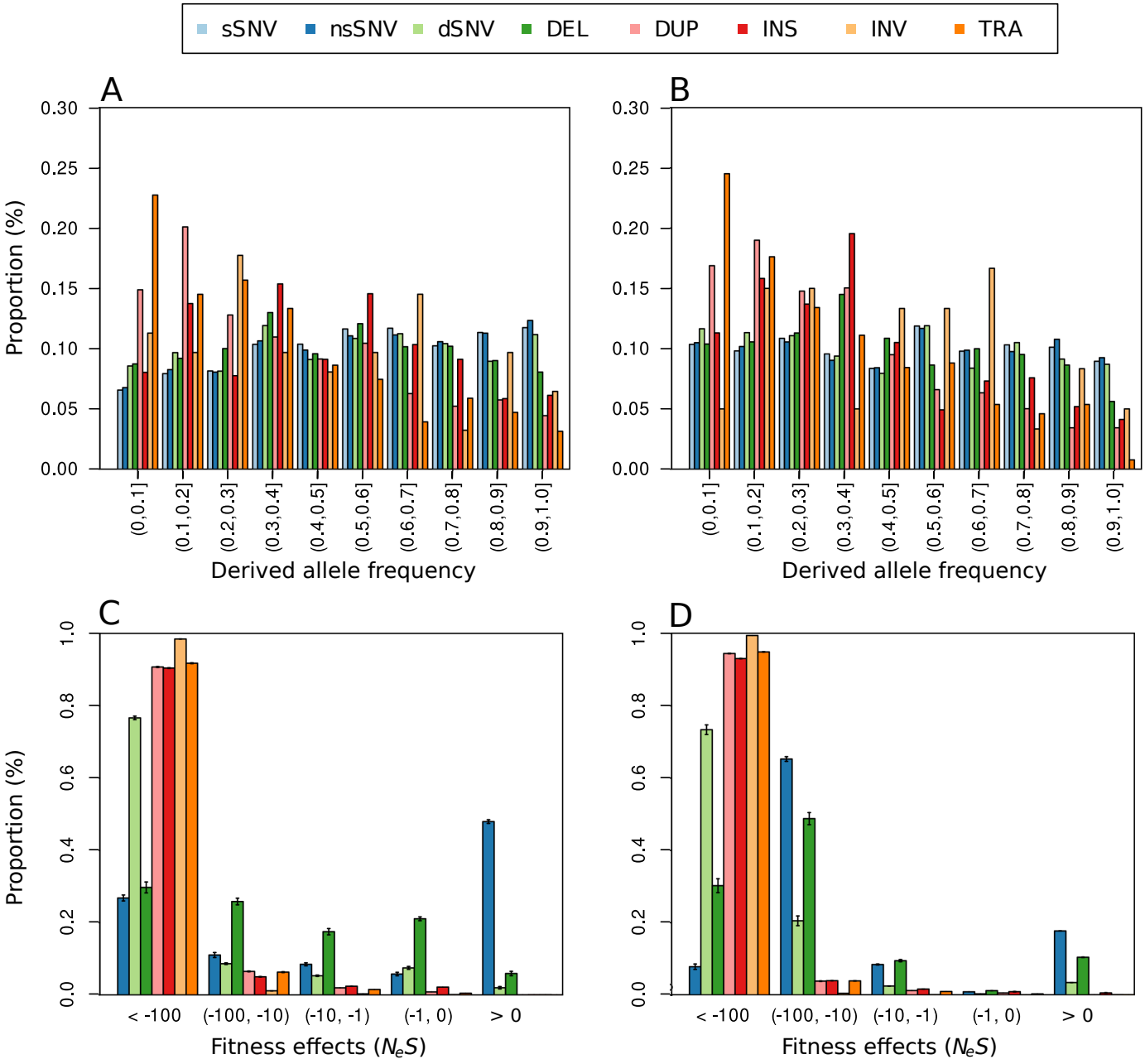

Fig. 7: Unfolded site frequency spectrum of deletions (DEL), duplications (DUP), insertions (INS), inversions (INV), and translocations (TRA) compared to synonymous single nucleotide variants (sSNV), non-synonymous SNV (nsSNV), and deleterious SNV (dSNV) for ten cultivars (A) and ten landraces (B) where three *Hordeum spontaneum* accessions were used as outgroup. Inferred distribution of fitness effects (N_eS) for the different types of SV and SNV in cultivars (C) and landraces (D) based on 100 bootstrap runs. Error bars indicate the 95% confidence interval.

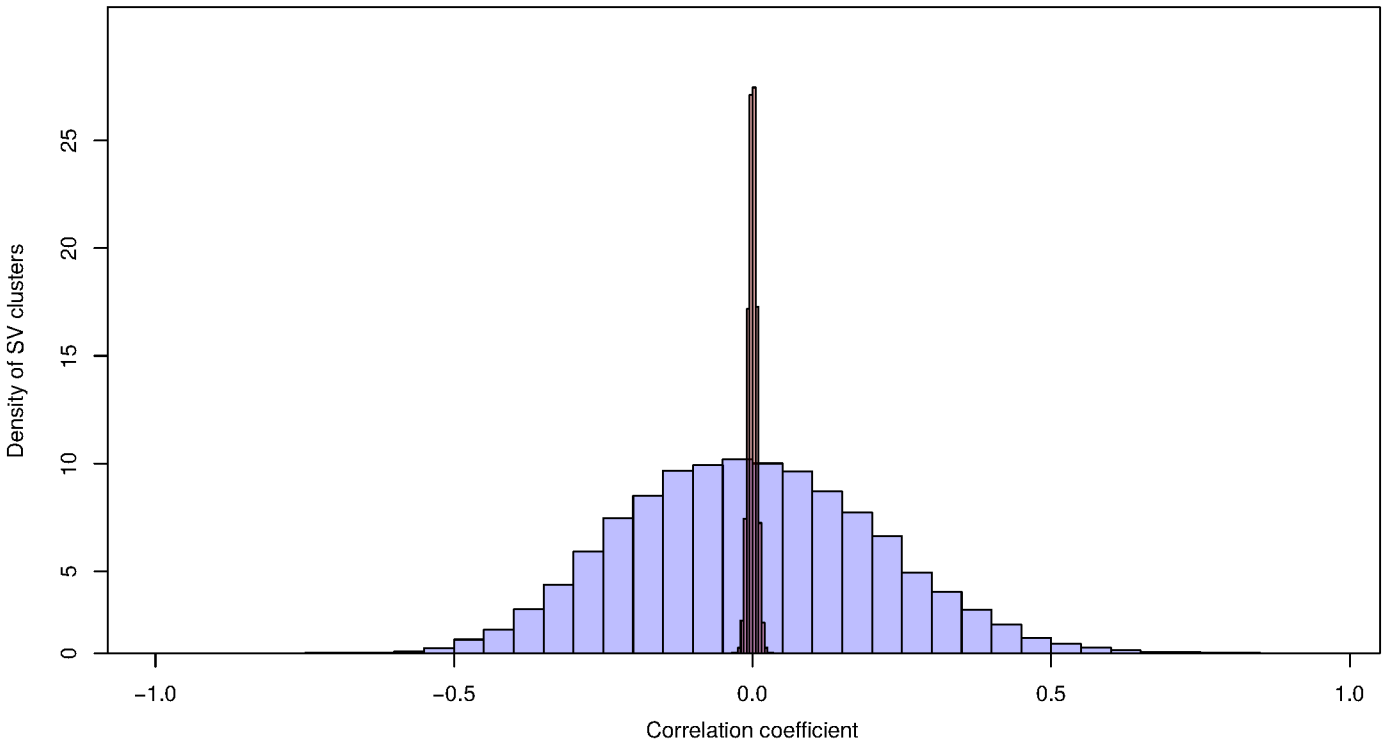

Fig. 8: Distribution of correlation coefficients of presence/absence pattern of all structural variant (SV) clusters (deletions, insertions, duplications, inversions) with minor allele frequency > 0.15 and the loadings of principal component 1 (19.7%) from a principal component analysis of gene expression data. The blue histogram shows the distribution for the detected SV clusters whereas the red histogram shows the distribution for random SV clusters with identical allele frequency.

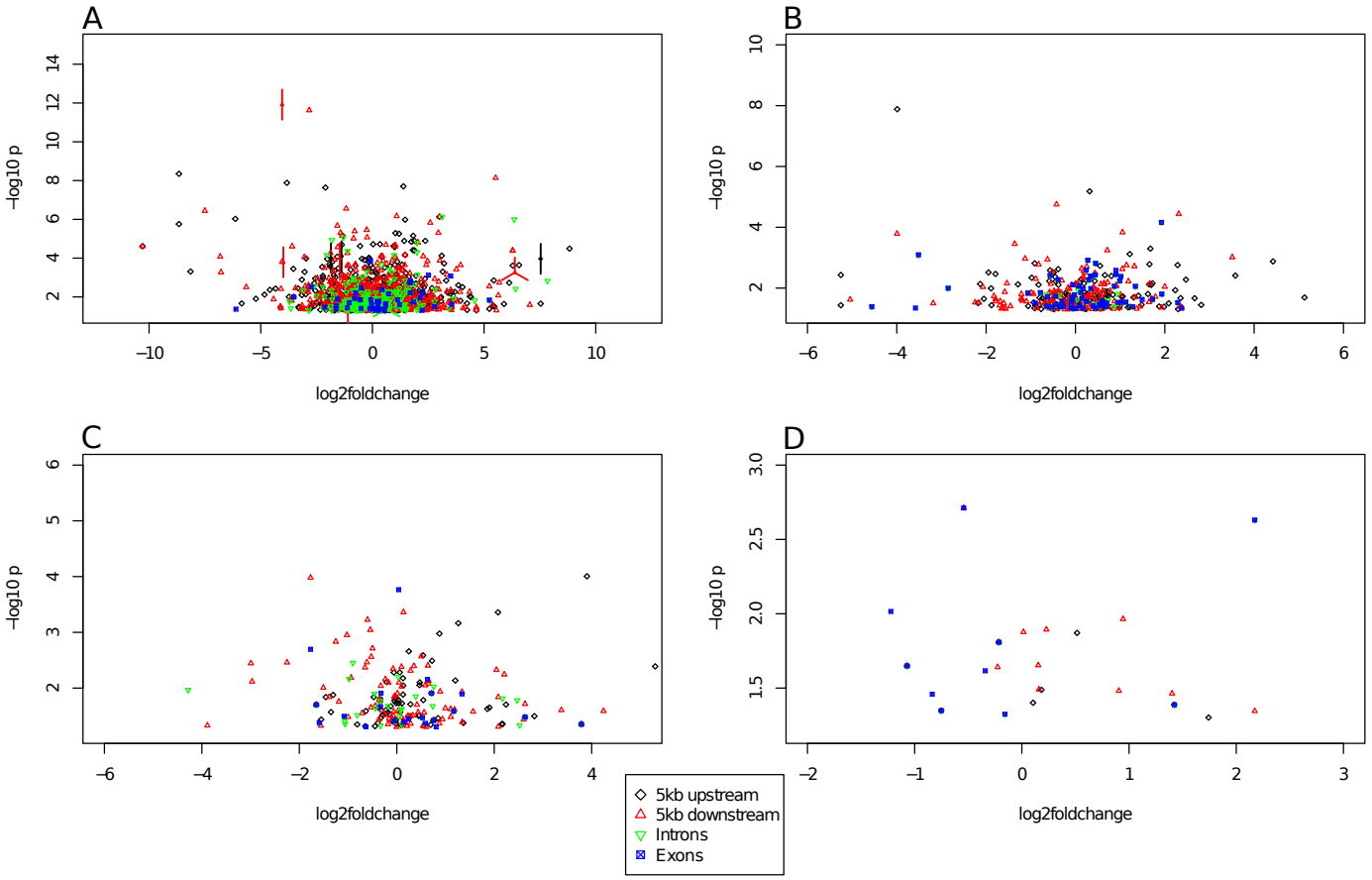

Fig. 9: Association of gene-associated (for details see Material & Methods) deletions (A), insertions (B), duplications (C), and inversions (D) with a minor allele frequency > 0.15 with the expression of individual genes assessed using the PK mixed linear model. The gene-associated structural variant (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.

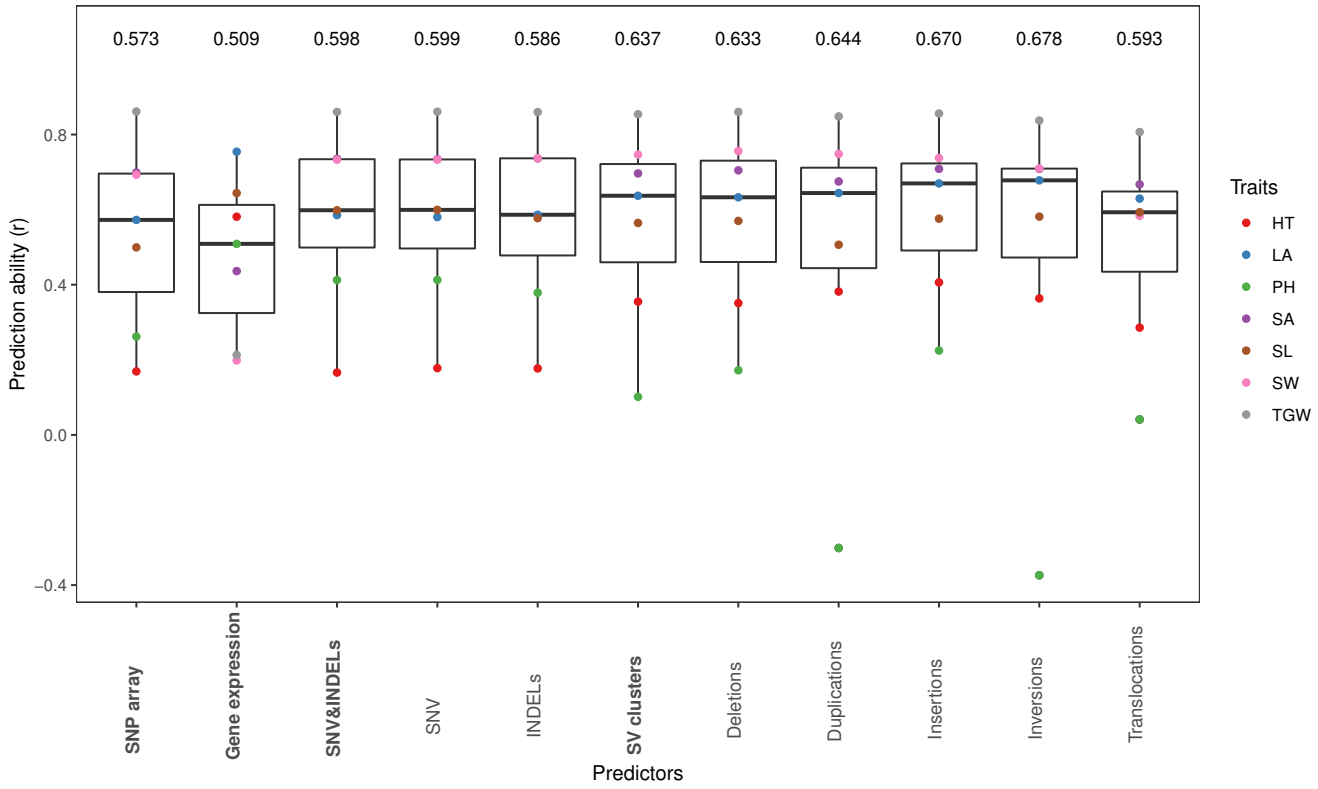

Fig. 10: Boxplot of the median prediction abilities across the seven traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), thousand grain weight (TGW) based on 23 inbreds using different predictors. The points in each box represent the medians 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 - 49bp, INDELs), as well as structural variant (SV) clusters individually as well as combined together.

SUPPLEMENTARY INFORMATION

Table S1: Predicted structural variants (SV) for PCR validation. Listed are all SV that were PCR validated including the names, sizes, primer positions, and the expected amplicon sizes. All sizes are given in bp.

	Primer position relative to SV start				Expected amplicon size	
SV names	left	right	2nd right	SV size	Morex	Unumli-Arpa
Del_A_1	-263	321		57	584	527
Del_A_2	-158	293		64	451	387
Del_A_3	-110	324		53	434	381
Del_A_4	-229	424		124	653	529
Del_A_5	-216	265		55	481	426
Del_A_6	-277	155		59	432	373
Ins_A_1	-167	243		57	353	410
Ins_A_2	-238	191		76	353	429
Ins_A_3	-234	258		91	401	492
Ins_A_4	-288	126		52	362	414
Ins_A_5	-266	239		57	448	505
Del_B_1	-391	2,704		1,937	3,095	1,158
Del_B_2	-891	1,699		1,303	2,590	1,287
Del_B_3	-462	4,446		4,144	4,908	764
Del_B_4	-374	3,687		2,940	4,061	1,121
Del_B_5	-797	2,529		2,263	3,326	1,063
Del_B_6	-459	2,273		1,393	2,732	1,339
Del_C_1	-364	316	11,313	10,778	680	899
Del_C_2	-103	280	5,692	5,355	383	440
Del_C_3	-231	375	28,406	27,937	606	700
Del_D_1	-262	120	287,036	286,558	382	740
Del_D_2	-361	371	91,956	91,411	732	906
Del_D_3	-248	224	54,918	54,481	472	685
Del_E_1	-169	348	460,621	460,240	517	550
Del_E_2	-279	239	405,578	405,029	518	828

Table S2: Inbred lines included in this study, their country of origin (CoO), row type, and year of release.

					Genome sequencing coverage		
Inbred name	BCC code	CoO	Row type	Year of release	seq	seq-trimmed	mapped
HOR1842	HOR1842	AFG	6	1935	27.4	26.6	25.7
HOR383	BCC1561	BGR	6	unknown	24.8	24.1	22.5
Sanalta	BCC929	CAN	2	1930	27.5	26.6	25.4
ItuNative	BCC502	CHN	6	unknown	23.6	23.0	21.3
Sissy	BCC1413	GER	2	1990	24.0	23.3	22.2
Georgie	BCC1381	GBR	2	1975	25.1	24.4	23.6
SprattArcher	BCC1415	GBR	2	1943	23.1	22.4	22.0
Lakhan	BCC533	IND	6	unknown	21.6	21.0	20.1
Kharsila	HOR11403	IND	6	before 1911	26.7	25.9	24.0
W23829/803911	HOR11374	ISR	2	unknown	23.6	23.0	22.2
Namhaebori	BCC667	KOR	6	unknown	22.3	22.0	21.5
IG128216	BCC118	LBY	6	1983	21.2	21.0	20.8
IG128104	BCC173	PAK	6	1974	23.8	23.2	22.3
K10693	BCC1491	RUS	6	unknown	21.0	20.5	19.8
IG31424	BCC190	SYR	2	1981	23.5	22.8	21.7
HOR12830	HOR12830	SYR	6	unknown	25.8	25.0	23.4
HOR7985	HOR7985	TUR	2	before 1969	23.3	22.5	21.9
K10877	BCC1503	TKM	6	unknown	25.5	24.8	23.7
HOR8160	HOR8160	TUR	2	before 1969	24.4	23.8	22.7
Ancap2	BCC807	URY	6	1950	27.0	26.2	24.5
CM67	BCC846	USA	6	1983	23.8	23.1	22.2
Kombyne	BCC893	USA	6	1975	21.5	20.9	20.0
Unumli-Arpa	BCC1470	UZB	2	unknown	23.5	22.9	21.8

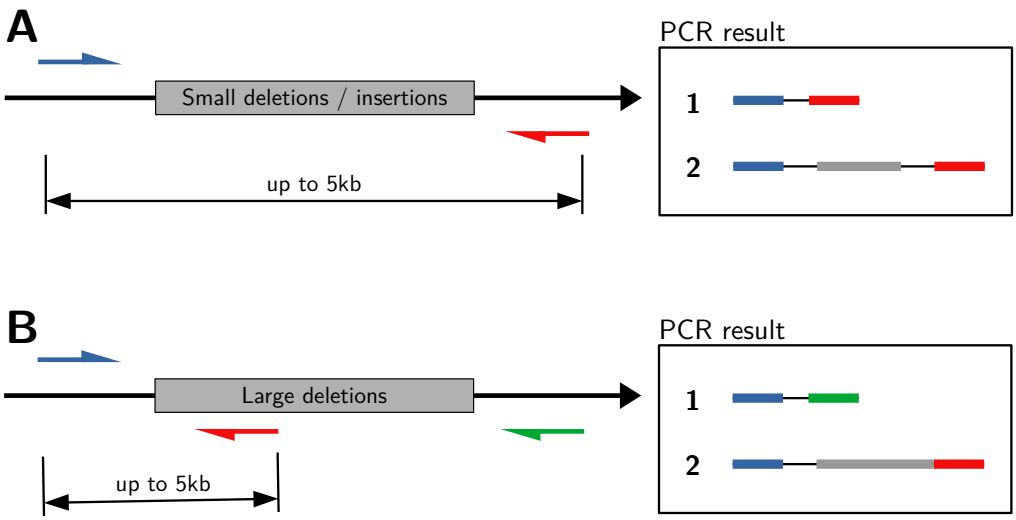

Fig. S1: Graphical illustration of the primer design strategy created to validate structural variant (SV) predictions in the reference genome Morex and Unumli-Arpa. The primer design strategy had to be adjusted depending on the size of the SV. Smaller deletions (A) and insertions (up to ∼5kb) were validated with a pair of two primers (blue/red arrow) flanking the SV (gray box). Larger deletions (B) were validated either by primer 1 (blue) and primer 2 (red) in case of presence or by primer 1 (blue) and primer 3 (green) in case of absence. The predicted PCR results, the absence (1) and presence (2) of the SV sequence in the PCR fragment, are shown on the right.

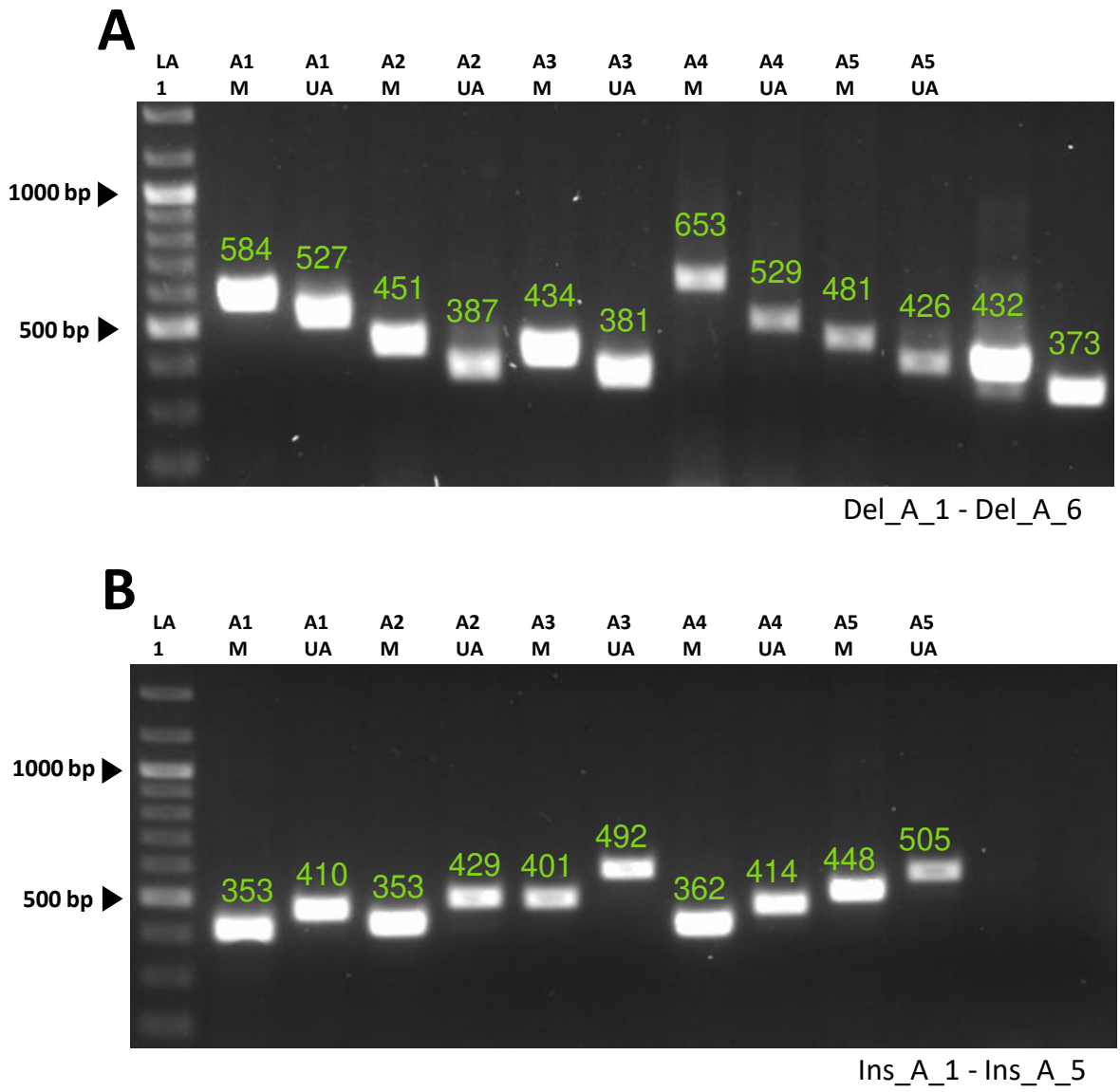

Fig. S2: PCR validation results for small structural variants (SV) as documented after the gel electrophoresis. PCR amplified fragments are shown separated by size for the reference genotype Morex (M) and the genotype Unumli-Arpa (UA). Predicted fragment size based on the SV predictions are illustrated by numbers. The numbers are colored based on the validation success. Fragment size agreement between PCR and prediction (green) or disagreement (red). Results are shown for six small deletions (A) and six small insertions (B) of the SV length category A (50 - 300bp). DNA ladder used: GeneRuler 100bp Plus, Thermo Fisher (LA 1).

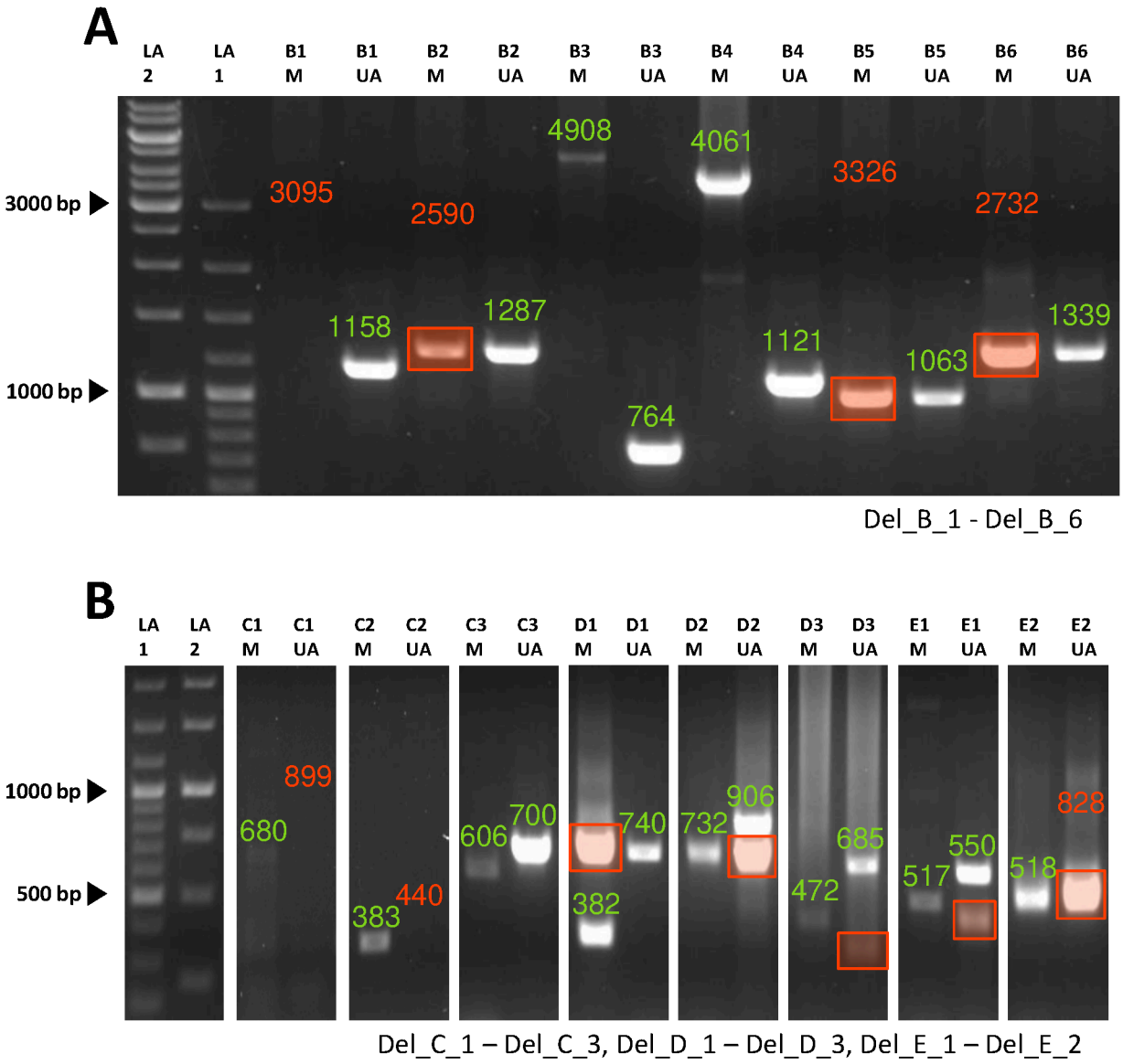

Fig. S3: PCR validation results for large structural variants (SV) as documented after the gel electrophoresis. PCR amplified fragments are shown separated by size for the reference genotype Morex (M) and the genotype Unumli-Arpa (UA). Predicted fragment size based on the SV predictions are illustrated by numbers. The numbers are colored based on the validation success. Fragment size agreement between PCR and prediction (green) or disagreement (red). Additional not predicted fragments are marked by a red box. Results are shown for six deletions of the SV length category B (0.3 - 5kb) (A) and 8 deletions of the SV length category C (5 - 50kb), D (50 - 250kb), and E (0.25 - 1Mb) (B). DNA ladder used: GeneRuler 100bp Plus (LA 1) and GeneRuler 1kb, Thermo Fisher (LA 2).

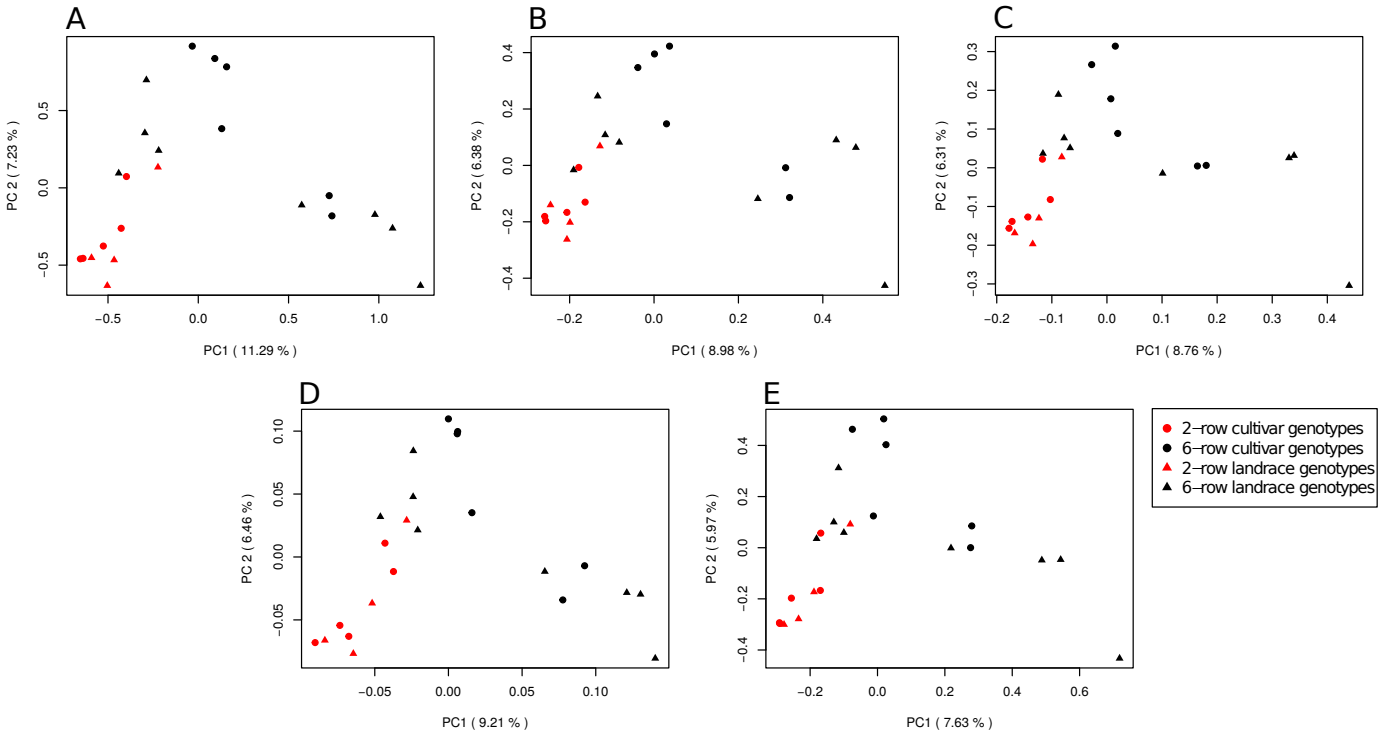

Fig. S4: Principal component analyses of the barley inbred lines considered in our study based on deletions (A), duplicatons (B), insertions (C), inversions (D), and translocations (E). PC 1 and PC 2 are the first and second principal component, respectively, and number in parentheses refer to the proportion of variance explained by the principal components. Symbols identify landrace and cultivar inbreds and colors their row number.

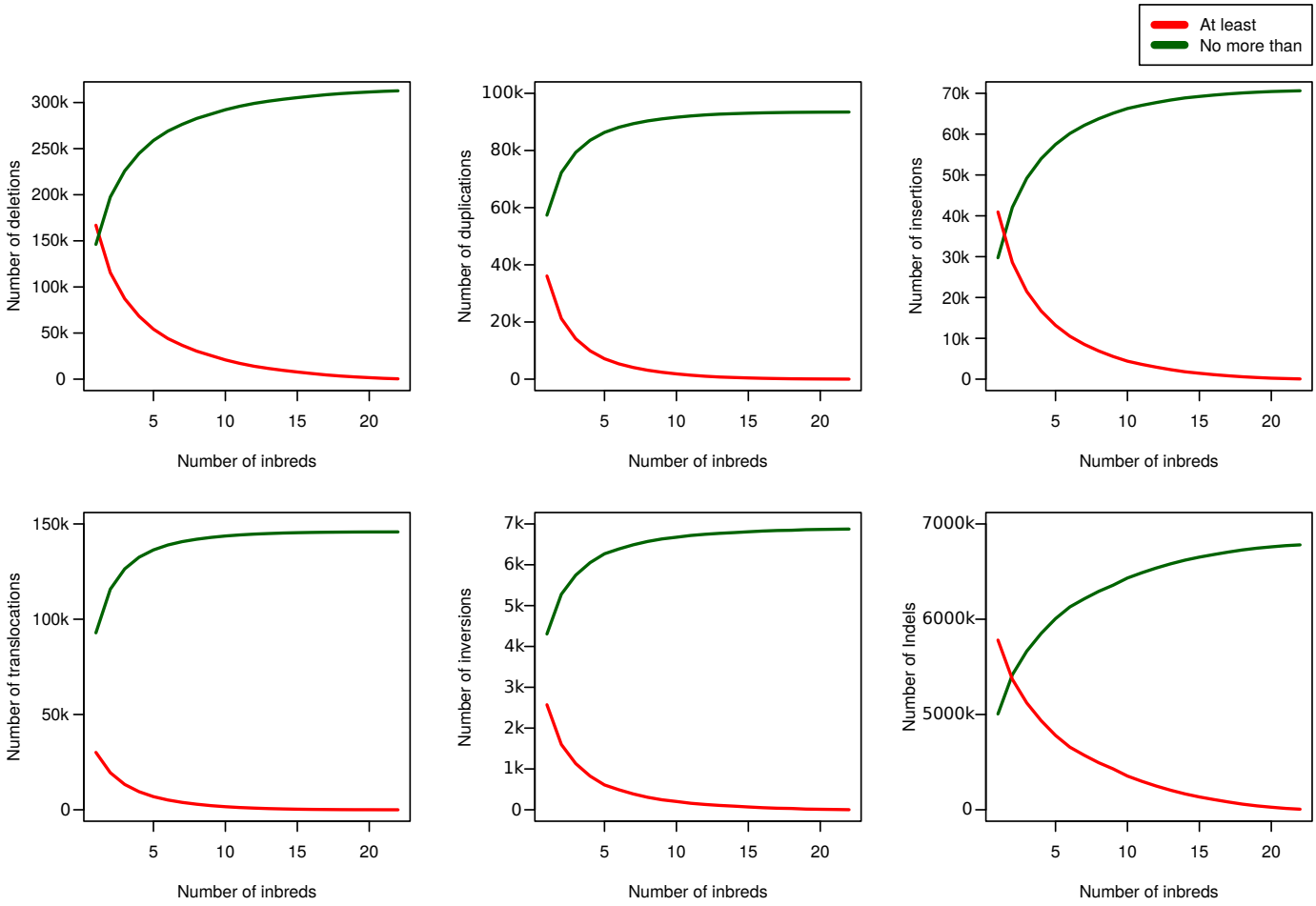

Fig. S5: Number of structural variant (SV) clusters for the different types of SV which were detected in at least (red) or no more than (green) the given number of inbreds.

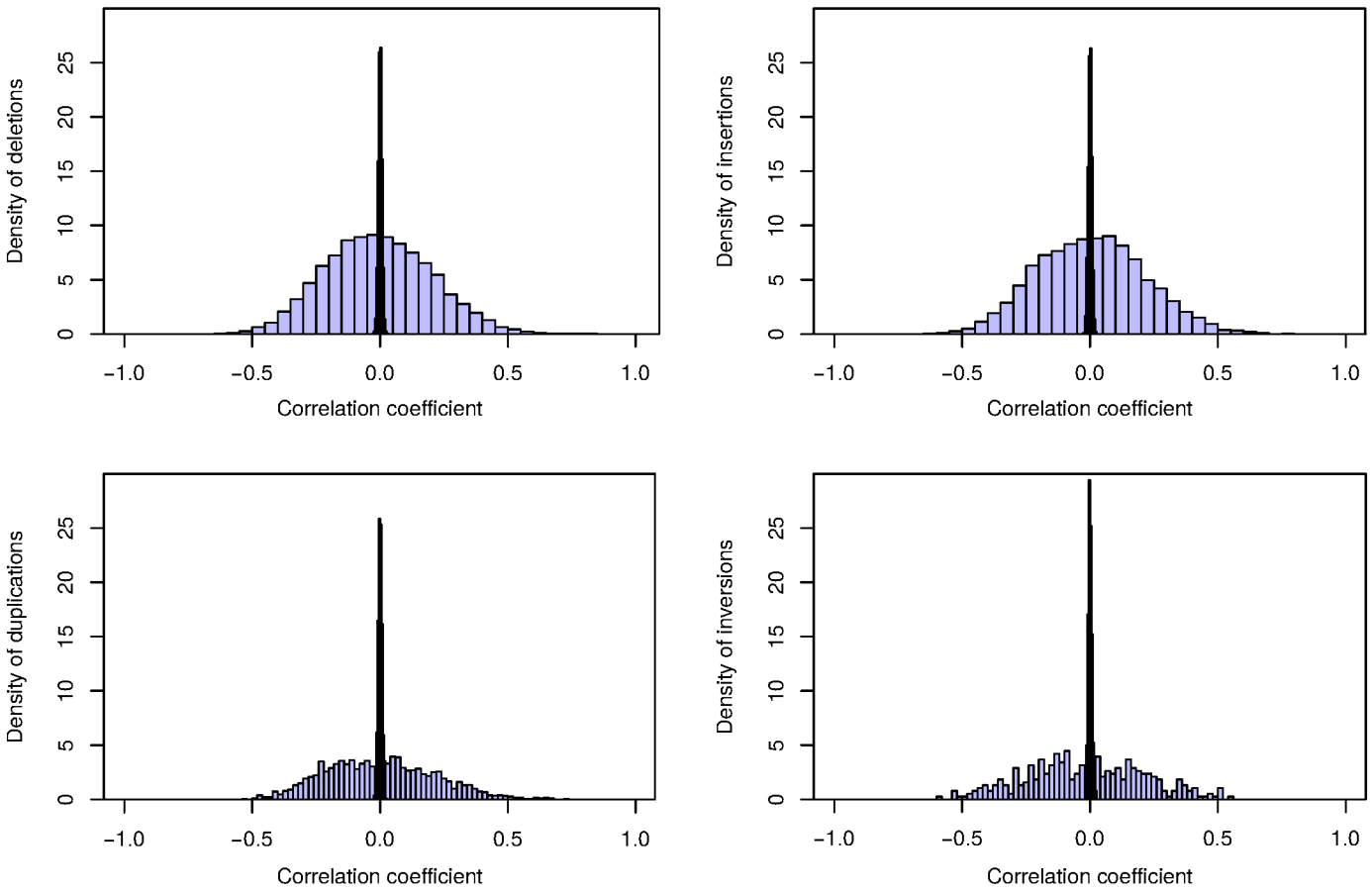

Fig. S6: Distribution of correlation coefficients of presence/absence pattern of deletions, insertions, duplications, and inversions with minor allele frequency > 0.15 and the loadings of principal component 1 (19.7 %) from a principal component analysis of gene expression data. The blue histogram shows the distribution for the detected SV clusters whereas the red histogram shows the distribution for random SV clusters with identical allele frequency.

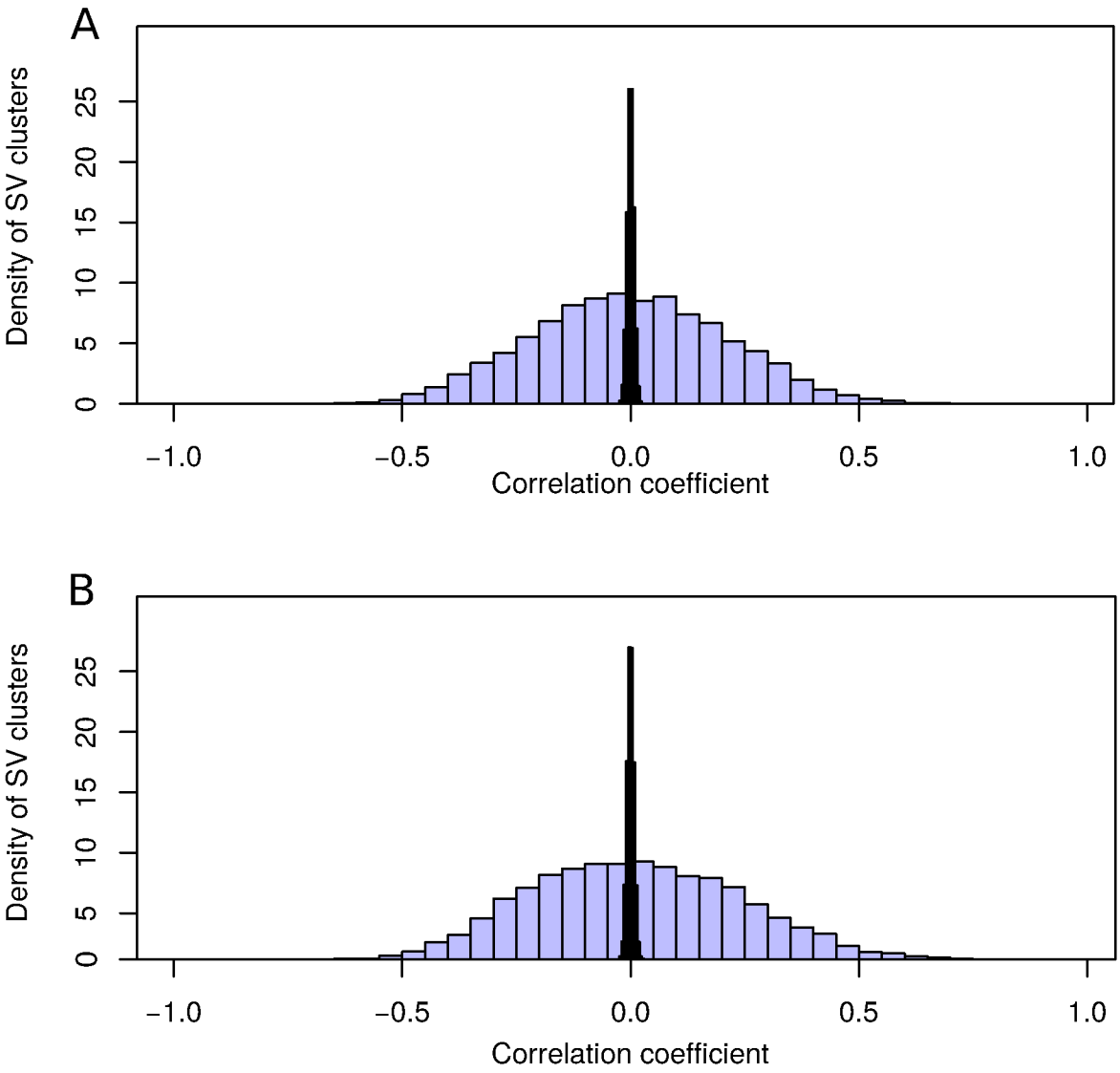

Fig. S7: Distribution of correlation coefficients of presence/absence pattern of SV clusters with minor allele frequency > 0.15 and the loadings of principal component 2 (8.2 %) (A), and 3 (7.1 %) (B) from a principal component analysis of gene expression data. The blue histogram shows the distribution for the detected SV clusters whereas the red histogram shows the distribution for random SV clusters with identical allele frequency.

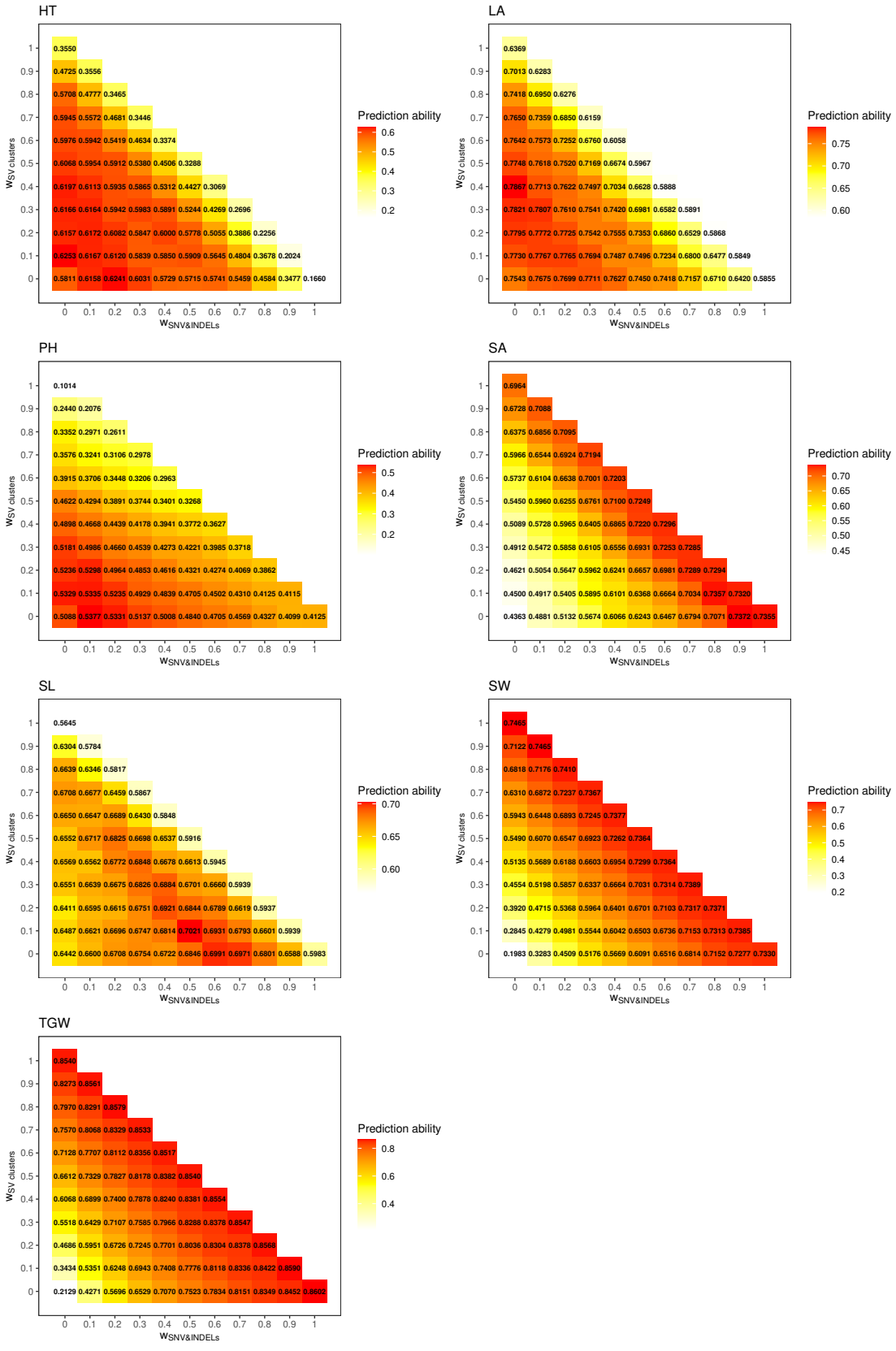

Fig. S8: Prediction ability for the seven phenotypic traits heading time (HT), leaf angle (LA), plant height (PH), seed area (SA), seed length (SL), seed width (SW), and thousand grain weight (TGW) from 23 inbreds for 66 combinations of the joined weighted matrices which differ in the weights of three predictors single nucleotide variants (SNV) and small insertions and deletions (2 - 49bp, INDELs, SNV&INDELs, x-axis), structural variant (SV) clusters (y-axis), and gene expression. Plotted values represent medians across 200 cross-validation runs.

8. List of publications

1. **Weisweiler, M.**[1], A. de Montaigu[1], D. Ries, M. Pfeifer, B. Stich. 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 20:787

2. Freire, R.[1], **M. Weisweiler**[1], R. Guerreiro[1], N. Baig, B. Hüttel, E. Obeng-Hinneh, J. Renner, St. Hartje, K. Muders, B. Truberg, A. Rosen, V. Prigge, J. Bruckmüller, J. Lübeck, B. Stich. 2021. Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. G3 Genes|Genomes|Genetics 11:jkab330

3. Casale, F. A., D. Van Inghelandt, **M. Weisweiler**, J. Li, B. Stich. 2021. Genomic prediction of the recombination rate variation in barley - a route to highly recombinogenic genotypes. Plant Biotechnology Journal https://doi.org/10.1111/pbi.13746

4. **Weisweiler, M.**, C. Arlt[1], P.-Y. Wu[1], D. Van Inghelandt, T. Hartwig, B. Stich. 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. PLOS Genetics In review

5. Wu P.-Y., B. Stich, **M. Weisweiler**, A. Shrestha, A. Erban, P. Westhoff, D. Van Inghelandt. 2022. Improvement of prediction ability by integrating multi-omic datasets in barley. BMC Genomics 23:200

6. **Weisweiler M.**, B. Stich. 2022. Benchmarking of structural variant detection in the tetraploid potato genome using linked-read sequencing. In preparation

[1]Contributed equally

9. Acknowledgements

I am very grateful to my academic supervisor Benjamin Stich for giving me the opportunity to do my PhD in his group and for his advice, suggestions, and support during this thesis work.
Thanks to Gunnar Klau for agreeing to act as co-supervisor for my thesis.
Special thanks go to my former colleagues David Ries and Amaury de Montaigu for introducing me to the field of next generation sequencing analyses during my master thesis. You are probably the reason for my specification in this research area and I am very thankful for your guidance in the beginning of my PhD.
Many thanks to Benjamin Stich, Amaury de Montaigu, David Ries, Mara Pfeifer, Ruth Freire, Ricardo Guerreiro, Nadia Baig, Delphine Van Inghelandt, Christopher Arlt, Po-Ya Wu, Thomas Hartwig, Bruno Hüttel, Evelyn Obeng-Hinneh, Juliane Renner, Stefanie Hartje, Katja Muders, Bernd Truberg, Arne Rosen, Vanessa Prigge, Julien Bruckmüller, and Jens Lübeck for being co-authors of the publications.
Thanks to Delphine Van Inghelandt, Benjamin Stich, Ines Sigge, Florian Esser, Michael Schneider, Asis Shrestha, Po-Ya Wu, Nadia Baig, Federico Casale, Ricardo Guerreiro, Ruth Freire, Mara Pfeifer, Christopher Arlt, Marius Kühl, Francesco Cosenza, Maria Schmidt, Yanrong Gao, Stephanie Krey, Vesna Lamesic, Anja Kyriacidis, Suresh Bonthala, Agata Stoltmann, George Alskief, David Ries, Amaury de Montaigu, and all unmentioned members and former members of the Institute for Quantitative Genetics and Genomics of Plants for creating a great work atmosphere.
The financial support from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and from the Federal Ministry of Food and Agriculture (Fachagentur Nachwachsende Rohstoffe) is gratefully acknowledged. Furthermore, I would like to thank the breeding companies Böhm-Nordkartoffel Agrarproduktion GmbH & Co. OHG, Nordring-Kartoffelzucht- und Vermehrungs- GmbH, and SaKa Pflanzenzucht GmbH & Co. KG for the good collaboration.
I would like to thank my family and friends for the interest in my work. Special thanks go to my wife Julia for her considerate support and for keeping me sane throughout the years.