ORIGINAL ARTICLE

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

Marius Weisweiler¹ · Christopher Arlt¹ · Po-Ya Wu¹ · Delphine Van Inghelandt¹ · Thomas Hartwig² · Benjamin Stich¹,³

Received: 11 May 2022 / Accepted: 3 August 2022 / Published online: 27 August 2022
© The Author(s) 2022

Abstract

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

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

Introduction

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

The development of sequencing by synthesis pioneered by Frederick Sanger (Sanger et al. 1977) enabled in the following years the first sequenced genomes of prokaryotes (e.g., Escherichia coli) and eukaryotes (e.g., yeast) (Goffeau et al. 1996; Blattner et al. 1997). Next milestones of sequencing by synthesis were the sequenced genomes of
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 Pacific Biosciences or Nanopore sequencing data which results in high costs if many accessions should be sequenced and, thus, is not affordable for many research groups. In contrast, short-read sequencing is well-established for SV detection in the human genome (Chaisson et al. 2019; Ebert et al. 2021). Various software tools have been developed to detect SV from short-read sequencing data and were benchmarked based on human genomes (Cameron et al. 2019; Kosugi et al. 2019).

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

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

In humans, SV have been described to have an up to ~50-fold 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; McCollan and Tabrizi 2018). In plants, individual SV have been associated with traits such as aluminum tolerance in maize (Maron et al. 2013), disease resistance and domestication in rice (Xu et al. 2012), or plant height (Li et al. 2012) and heading date (Nishida et al. 2013) in wheat. In barley, individual SV have been associated with traits such as Boron toxicity tolerance (Sutton et al. 2007) and disease resistance (Muñoz-Amatriain et al. 2013). In grapevine and rice, it has been shown that SV have a low variant frequency due to purifying selection (Zhou et al. 2019; Kou et al. 2020). However, few studies have examined the ability to predict quantitatively inherited phenotypic traits using SV in comparison to SNV.

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

Methods

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

Computer simulations

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

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

\[
\text{Sensitivity} = \frac{TP}{(TP + FN)} \\
\text{Precision} = \frac{TP}{(TP + FP)} \\
F1 - core = 2 \times \frac{(\text{Precision} \times \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})}
\]

with a logical and (“&”). If the precision of the combination of the second step increased to \(\geq 95\%\) in all SV length categories, SV calls of this combination were kept for the particular SV type and were combined with a logical or with those of the first step. The threshold of \(\geq 95\%\) 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 S6) using the MSTRAT algorithm (Gouesnard 2001). These inbreds are the parents of the double round robin population (Casale et al. 2022). Paired-end sequencing libraries with an insert size of 425 bp were sequenced (2x150 bp) to a \(\sim 25x\) coverage on the Illumina HiSeqX platform by Novogene Corporation Inc. (Sacramento, USA).

SV, INDELs, and SNV detection

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

Based on the results of the benchmarking of different SV callers using simulated data, the results of specific SV callers were combined as explained above. The final set of deletions for each inbred were those that were identified by Manta | GRIDSS | Pindel | Delly | (Lumpy & NGSEP) where homozygous-reference (0/0) and heterozygous variant (0/1) calls were removed. Additionally, deletions annotated by Pindel as RPL were removed. In analogy, the duplications were identified by Manta | GRIDSS | Pindel | (Delly & Lumpy). Insertions of the SV length category A were identified by Manta | GRIDSS | Delly, where insertions of the SV length categories B-E were called using Manta. Inversions were identified by Manta | GRIDSS | Pindel. Translocations were called from pairs of break points identified by Manta | GRIDSS | (Delly & Lumpy). INDELs were detected by GATK’s HaplotypeCaller | GRIDSS | Pindel where homozygous-reference (0/0) and heterozygous variant (0/1) calls were discarded. SV which were located in a region of the reference sequence, where the sequence only consists of N’s, were excluded. For genome regions, where break points of different SV overlapped or were inconsistent in the same inbred, only the smallest SV was considered.
The number of false positives could be increased by detecting large SV clusters; therefore, SV clusters larger than 1 Mb were not considered in our study. The SV of the 23 inbreds were grouped together to SV clusters based on the similarity of sizes and the position in the genome according to the following procedure. The distance from a SV to the next SV in such a SV cluster had to be smaller than 20 bp for the SV length category A and 50 bp for the SV length category B - E and the difference of the two break points had to be smaller than 10 or 50 bp as described above. SV with a larger difference between break points were kept as separate SV and SV clustering was pursuing. Each SV cluster was genotyped across the examined 23 barley inbreds.

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

**PCR validation of SV**

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

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

**Location of SV clusters**

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

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

SV hotspots were identified using the following procedure: The average number of SV clusters in non-overlapping 1 Mb windows across each of the seven chromosomes was determined. Using this number, we calculated for each window based on the Poisson distribution the expected number of SV clusters. Windows with more SV clusters than the Q99 of the expected Poisson distribution were designated as SV hotspots (Guan et al. 2021).

**Population genetic analyses**

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

**SV clusters and gene expression**

SV clusters which were assigned into one of the gene-associated SV categories, namely 5 kb up- or downstream, introns, and exons, were associated with the genome-wide gene expression of the 23 barley inbreds. Gene expression for the
seedling tissue measured as fragments per kilobase of exon model per million fragments mapped was available for all inbreds from an earlier study (Weisweiler et al. 2019). This information was the basis of a principal component analysis. For all gene-associated SV clusters with a minor allele frequency (MAF) > 0.15, Pearson’s correlation coefficient with the first three principal components was estimated, where the presence and absence of SV clusters were used as metric character. This analysis was performed to examine the association between SV clusters and genome-wide gene expression (Liu et al. 2020b). A permutation procedure with 1,000 iterations was used to test the mean absolute values of the correlations for their significance. In addition to this evaluation of the effect of SV clusters on the genome-wide gene expression level, we also examined the significance of the effect of gene-associated SV clusters with a MAF > 0.15 on the expression of individual genes. In order to do so, the mixed linear model with population structure and kinship matrix (PK model) (Stich et al. 2008) was used. The population structure matrix consisted of the first two principal components calculated from 133,566 SNV and INDELs derived from mRNA sequencing (Weisweiler et al. 2019). From the same information, the kinship matrix was calculated as described by Endelman and Jannink (2012).

**Assessment of phenotypic traits**

For the assessment of phenotypic traits under field conditions, the 23 inbreds were planted as replicated checks in an experiment laid out as an augmented row-column design. The experiment was performed in seven environments (Cologne from 2017 to 2019, Mechernich and Quedlinburg from 2018 to 2019) in Germany in which the checks were replicated multiple times per environment. For each environment, seven phenotypic traits were assessed. Heading time (HT) was recorded as days after planting, leaf angle (LA) was scored on a scale from 1 (erect) to 9 (very flat) on four-week-old plants, and plant height (PH, cm) was measured after heading in Cologne and Mechernich. Seed area (SA, mm²), 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 Sensors, Neubrandenburg, Germany).

**Prediction of phenotypes**

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

\[ y_{ijk} = \mu + E_j + G_i + (G \times E)_{ij} + \epsilon_{ijk}, \]  

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

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

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

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

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

**Precision and sensitivity of SV callers**

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

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

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

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

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

In a next step, different sequencing coverages from 1.5x to 65x were simulated and the performance of the best combination of SV callers for each of the SV types was compared to their performance with 25x sequencing coverage (Supplementary Fig. S1 ). For deletions, the F1-score, which is harmonic mean of the precision and sensitivity, for 65x sequencing coverage was ∼2% higher than for 25x sequencing coverage. Only marginal differences were observed between the F1-score of 65x or 25x sequencing coverage for calling duplications and inversions. Interestingly, the F1-score for calling translocations and insertions was with 2% and 9%, respectively, higher in the scenario with 25x than with 65x sequencing coverage. For 12.5x sequencing coverage, the F1-score was still on an high level with values > 80% for each SV type (Supplementary Fig. S2).

With a further reduced sequencing coverage, the F1-score also decreased. Finally, the performance of our pipeline to detect SV was evaluated based on 14x and 25x linked-read sequencing data. For all SV types and SV length categories, with the exception of deletions and duplications in SV length category D and A, respectively, the F1-score was 2–7% higher based on Illumina sequencing data than based on linked-read sequencing data.

---

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

| SV caller | Detection principle | Deletion | Insertion ≤500bp | Insertion >500bp | Inversion | Duplication | Translocation |
|-----------|---------------------|----------|-----------------|-----------------|-----------|-------------|--------------|
| Pindel    | x                   | x        | x               | x               | x         | x           |              |
| Delly     | x                   | x        | x               | x               | x         | x           |              |
| Lumpy     | x                   | x        | x               | x               | x         | x           |              |
| Manta     | x                   | x        | x               | x               | x         | x           | x            |
| GRIDSS    | x                   | x        | x               | x               | x         | x           | x            |
| NGSEP     | x                   | x        | x               | x               | x         | x           | x            |

1Ye et al. (2009), 2Rausch et al. (2012), 3Layer et al. (2014), 4Chen et al. (2016), 5Cameron et al. (2017), 6Duitama et al. (2014)
SV clusters across the 23 parental inbreds of the double round robin population

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

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

| SV caller | SV length category | Deletions | Insertions | Duplications | Inversions |
|-----------|--------------------|-----------|------------|--------------|------------|
|           |                    | A         | B          | C            | D          |
| 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 |

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

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

<sup>1</sup>Because of missing endpoint information no reciprocal overlap criterion applied
across the seven chromosomes. The proportion of SV clusters which were annotated as transposable elements varied from 1.4% for inversions to 51.5% for translocations.

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

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

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

![Fig. 1 Stacked bar graph of the number of different types of structural variant (SV) clusters detected in the 23 inbreds (A) and SV clusters which were detected in at least the given number of the inbreds (B)](image-url)
largest fraction were that for SV clusters of the SV length category C (55.9%). The average MAF of the individual SV types was the highest for insertions with 0.17, followed by deletions, inversions, translocations, and duplications with values of 0.14, 0.11, 0.10, and 0.10, respectively.

**Characterization of the SV clusters**

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

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

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

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

**Association of SV clusters with gene expression**

We tested if the SV clusters could be associated with the genome-wide gene expression differences of the 23 inbreds. As a first step, a principal component analysis of the gene expression matrix, which included all genes and inbreds, was performed. The loadings of all 23 inbreds on principal component (PC) 1 explained 19.7% of the gene expression variation and were correlated with the presence/absence status of all inbreds for each gene-associated SV cluster. The average absolute correlation coefficient of gene-associated SV clusters and the PC1 of gene expression was 0.17 and higher than the \(Q_{0.01}\) of the coefficient observed for randomized presence/absence pattern and the PC1 (Supplementary Fig. S10, Supplementary Fig. S11). Similar observations were made for the association of gene-associated SV clusters with PC2 and PC3 of 0.17 and 0.19, respectively, for the above-mentioned gene expression matrix (Supplementary Fig. S12). In addition, we investigated a possible association between SV clusters and gene expression on the basis of individual genes. For a total of 1,976 out of 21,140 gene-associated SV clusters a significant \((P < 0.05)\) association with the gene expression of the associated gene was observed (Fig. 5).

**Prediction of phenotypic variation from SV clusters**

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

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

**Fig. 3** The occurrence of deletions (A), insertions (B), duplications (C), inversions (D), small insertions and deletions (2–0.49 bp, INDELs, E), and single-nucleotide variants (SNV) (F) in five genomic regions

**Limitations of our study**

In our study, the different SV types were always determined in comparison against one reference sequence. The number of insertions present in this reference inbred determines the number of detected deletions and vice versa. However, this
is just a matter of nomenclature. Additionally, the usage of short-read sequencing data and only one reference sequence could lead to detect false positive SV calls, due to differences in the mapping efficiency of the evaluated inbreds due to differences in relatedness. In our study, however, the average mapping quality for the 23 inbreds was high and varied only moderately between 41 and 46. Therefore, the influence of the relatedness should be weak. However, this aspect should be considered when interpreting the SV data set.

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

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

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

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

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

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

![Fig. 5](image-url):

**Fig. 5** Association of gene-associated (for details see Material & Methods) deletions (A), insertions (B), duplications (C), and inversions (D) with a minor allele frequency > 0.15 with the expression of individual genes assessed using the PK mixed linear model. The gene-associated structural variant (SV) clusters were classified based on their occurrence relative to genes in 5kb up- or downstream, introns, and exons. Values of SV clusters with the same coordinates are illustrated as points with edges, where each edge represents one SV cluster.
recommended. This approach was then used for SV detection in the set of 23 spring barley inbreds.

Validation of SV in the barley genome

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

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

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

Characteristics of SV clusters in the barley gene pool

Across the 23 spring barley inbreds that have been selected out of a world-wide diversity set to maximize phenotypic and genotypic diversity (Weisweiler et al. 2019), we have identified 458,671 SV clusters (Table 3). This corresponds to 1 SV cluster every 9,149 bp and corresponds to what was observed by Jayakodi et al. (2020). This number is in agreement with the number of SV clusters detected for cucumber (9,788 bp−1) (Zhang et al. 2015) or peach (8,621 bp−1) (Guan et al. 2021). Other studies have revealed a higher number of SV clusters than observed in our study. This might be due to the considerably higher number of re-sequenced accessions for each trait. The predictors were: features from SNP array, gene expression, single nucleotide variants (SNV) and small insertions and deletions (2–49 bp, INDELs), as well as structural variant (SV) clusters individually as well as combined together.
in rice (214 bp⁻¹) (Fuentes et al. 2019), tomato (3,291 bp⁻¹) (Alonge et al. 2020), and grapevine (1,260 bp⁻¹) (Zhou et al. 2019).

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

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

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

**Association of SV clusters with transcript abundance**

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

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

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

**Genomic prediction**

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

We observed that the ability to predict these seven traits was higher for SV clusters compared to the benchmark data from a SNP array (Fig. 6). This might be explained by the considerably higher number of SV clusters than variants included in the SNP array. However, we observed the same trend when comparing the prediction ability of SV clusters to that of the much more abundant SNV & INDELs. This indicates that the SV clusters comprise genetic information that is not comprised by SNV & INDELs. Our result is supported by the observation that when examining the combination of SNV and INDELs with gene expression and SV clusters to predict phenotypic traits, an increase of the prediction ability was observed compared to the ability observed for the individual predictors (Supplementary Table S5). Furthermore, our observation of a different prediction ability...
between SV clusters and SNV & INDELs can be explained by a lower extent of LD between SV clusters and linked SNV compared to that between SNV and linked SNV (Supplementary Table S4). These findings together illustrate the high potential of using SV clusters for the prediction of phenotypes in diverse germplasm sets. Such type of applications might be used also in commercial plant breeding programs. From a cost perspective such approaches will be realistic if SV detection is possible from low coverage sequencing. This might be possible when comprehensive reference sets of SV per species are available as was, for example, generated in our study for barley. However, this requires further research.

**Usefulness of SV information for QTL fine mapping and cloning**

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

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

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

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

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

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

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

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

**Declarations**

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

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

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

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

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

Altmuhler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flice P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander
ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVeyan GA et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

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

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

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Baker M (2012) Structural variation: the genome’s hidden architecture. Nat Methods 9:133–137

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimming tool 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

Casadevall M, Downie J, Furlong E, Klok A, Sachdev K, van der Loo J (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27:1–11

Casale F, Van Inghelandt D, Stockwell T, Turner R, Venter E, Wang M, Wen W, Wu M, Xie M, Ziadieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

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

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

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

Dujon B, Winge RR, Mauleon R, Alexandrov N (2019) Structural variants in 3000 rice elite variety. G3 Genes Genomes Genetics 11:jkab330

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

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

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

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

Goffea A, Barrett BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippen P, Tettelin H, Oliver SG (1996) Life sciences 135:3511–3529

Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puris V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Robmlad D, Ruhlfe B, Scott R, Siter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Ni Tint N, Tse V, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolandö KV, Karikas B, Karijaril A, Mi H, Lazareva B, Hatton T, Nar- echanina A, Diemer K, Muruganjanu A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Bick L, Caminina M, Carnes-Stine J, Cauk P, Chiang HY, Coyne M, Dahlke C, Deslatelles M, Aomsbork M, Donnelly M, Ely D, Esperham S, Fosler C, Gire H, Gianowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan K, Kajsa J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen W, Wu D, Wu M, Xie M, Ziadieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Craig-Holmes AP, Moore FB, Shaw MW (1973) Polymorphism of human C-band heterochromatin. I. Frequency of variants. Am J Hum Genet 25:181–192
Gong T, Hayes VM, Chan EK (2021) Detection of somatic structural variants from short-read next-generation sequencing data. Brief bioinform 22:1–15

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

Guan J, Xu Y, Yu F, 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, Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, Zhao J, Jiang Q (2021) Gouesnard B (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

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

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

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

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

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kania K, Wilm M, Fiebig A, Budak H, Xu D et al (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289

Kovalchuk O, Tauscher J, Cattanach BM, Toonen IR, Malitsky S, Colman J, Stoye J, Volff JN (2012) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84

Lee B, Yoon S, Cho H, Lee W, Lim B, Zeng Z, Kim J, Hwang J, Park H, Lee S, Rhee J, Seong J, Kang J, Yoon S, Lee J, Shin D, Lee J (2020) Whole-genome sequencing and structural variant discovery in Burkitt lymphomas. Nature 587:284–289

S, Keivanfar N, Kyriazopoulou-Panagiotopoulou S, Lek M et al (2019) Resolving the full spectrum of human genome variation using linked-reads. Genome Res 29:635–645

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

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radvuch V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastei A, Simková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriain M, Ounit R, Wannamaker S, Bolser D, Colmsee C, Schmutzer T, Alyiev-Schnorr L, Grasso S, Tanskanen J, Chailanya A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, Mccooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisijuk L, Houben A, Dolesjel A, Aylings S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Cacciato M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–437

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

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

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

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

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

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

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

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

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

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

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Schüle B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, Liu M, Fuentes RR, Talag J, Wang J, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Liu W, Liu L, Jackson D, Jan Y (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51:1052–1059

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

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

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.