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

Marius Weisweiler, Benjamin Stich

Genomics 115 (2023) 110568

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

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

Marius Weisweiler [a], Benjamin Stich [a,b,c,*]

[a] Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany
[b] Cluster of Excellence on Plant Sciences, From Complex Traits towards Synthetic Modules, Universitätsstraße 1, 40225 Düsseldorf, Germany
[c] Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, 50829 Köln, Germany

ARTICLE INFO

Keywords:
Structural variants
Potato
Tetraploidy
Linked-read sequencing
Plant genomics

ABSTRACT

It has recently been shown that structural variants (SV) can have a higher impact on gene expression variation compared to single nucleotide variants (SNV) in different plant species. Additionally, SV were associated with phenotypic variation in several crops. However, compared to the established SV detection based on short-read sequencing, less approaches were described for linked-read based SV calling. We therefore evaluated the performance of six linked-read SV callers compared to an established short-read SV caller based on simulated linked-reads in tetraploid potato. 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. We observed high break point resolutions (BPR) detecting short SV and slightly lower BPR for large SV. Our observations highlighted the importance of short-read signals provided by Manta and LinkedSV to detect short SV. Manta and NAIBR performed well for detecting larger deletions, inversions, and duplications. Detected large SV were weakly influenced by the HI. Furthermore, we illustrated that large insertions can be assembled by Novel-X. Our results suggest the usage of the short-read and linked-read SV callers Manta, NAIBR, LinkedSV, and Novel-X based on at least 90x linked-read sequencing coverage to ensure the detection of a broad range of SV in the tetraploid potato genome.

1. Introduction

Structural variants (SV) are commonly defined as genomic rearrangements between individuals or haplotypes that are larger than 49 bp [19]. SV can occur as deletions, insertions, duplications, inversions, or translocations in the genome. SV were more strongly associated with gene expression variation compared to single nucleotide variants (SNV) in human [9] and were also associated with transcript abundance in crops such as maize and tomato [2,57]. Additionally, SV were associated with phenotypic variation in several plant species such as wheat and rice [32,39,56]. In potato, copy number variation at a limited number of loci was associated with the level of gene expression [23].

Due to the technical improvements of DNA sequencing and novel algorithms [19], 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 [4,29] and was also evaluated and used recently for plant genomes [16,17]. However, the reliable detection of SV based on short-read sequencing is challenging due to the necessity of confidently mapped read-pairs [14]. Additionally, repetitive regions are associated with the occurrence of SV [21], where split and paired-end reads can have a low mapping quality due to multi-mapping [14]. These issues can be avoided by using long-read sequencing [11]. However, this approach in turn is associated with high costs and therefore, it is less efficient in breeding-related applications.

Recently, linked-read sequencing was proposed [50,51]. For linked-read sequencing, paired-end short reads are derived from 50 to 100 kb DNA molecules [12], which is at least as long as the read length of most long-read sequencing approaches (cf. [53]). 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

* Corresponding author at: Institute for Quantitative Genetics and Genomics of Plants, Universitätsstraße 1, 40225 Düsseldorf, Germany.
E-mail address: benjamin.stich@hhu.de (B. Stich).

https://doi.org/10.1016/j.ygeno.2023.110568
Received 28 October 2022; Received in revised form 12 January 2023; Accepted 18 January 2023
Available online 23 January 2023

regions in the genome is very low [12]. Therewith, linked-read sequencing provides long-range information as long-read sequencing [19] and has the advantages of a high accuracy and low costs as short-read sequencing [51]. However, compared to the established SV detection based on short-read sequencing, less approaches have been described and evaluated for linked-read based SV calling.

Eight linked-read SV callers were described until today, namely LongRanger [58], GROC-SVs [46], NAIBR [12], ZoomX [55], LinkedSV [14], Novel-X [35], VALOR2 [24], and LEVIATHAN [37]. 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 short deletions. In addition, this tool uses fragments with shared barcodes between two genomic locations and enriched fragment endpoints near break points to detect larger SV [14]. 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 [19]).

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 [37]. 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, no earlier study evaluated SV calling for an autotetraploid genome and which examined the effect of the haplotype incidence (HI) on SV detection. Additionally, the detection of SV in the tetraploid potato genome is of high interest, due to the potential usage of SV as genetic markers in genome-wide association studies [36] or genomic selection [47,54] to increase the gain of selection in this important crop species.

Therefore, the objectives of our study were to i) compare the performance of SV callers based on linked-read sequencing to that of short-read sequencing, ii) examine the influence of SV type, SV length, 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.

2. Material and methods

2.1. Simulation preparation and genome mutation

We used Mutation-Simulator (version 2.0.3) [30] to simulate deletions, duplications, inversions, and insertions in the first and second chromosome of the dAg1_v1.0 potato reference sequence [15] 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 $7.0\text{x}10^{-6}$ (~ 800-1000 SV) for the SV length categories A - C, $7.0\text{x}10^{-7}$ (~ 90 SV) for D, and $3.5\text{x}10^{-7}$ (~ 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. 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 short 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 (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.

2.2. Linked-read simulation and mapping

LRSim (version 1.0) [33] was used to simulate linked-reads with the following parameters (-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. [33] for *Arabidopsis thaliana* which have a similar genome size as the first two chromosomes of the dAg1_v1.0 reference sequence [15]. Linked-reads were mapped against the non-mutated dAg1_v1.0 reference sequence with LongRanger wgs (version 2.2.2).

2.3. SV calling and filtering

LRez (version 2.2.2) [38] 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) [8] 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) [35], and NAIBR [12] were evaluated as linked-read SV callers (Table 1). Additionally, LinkedSV and LongRanger can detect short 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) [28] 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.

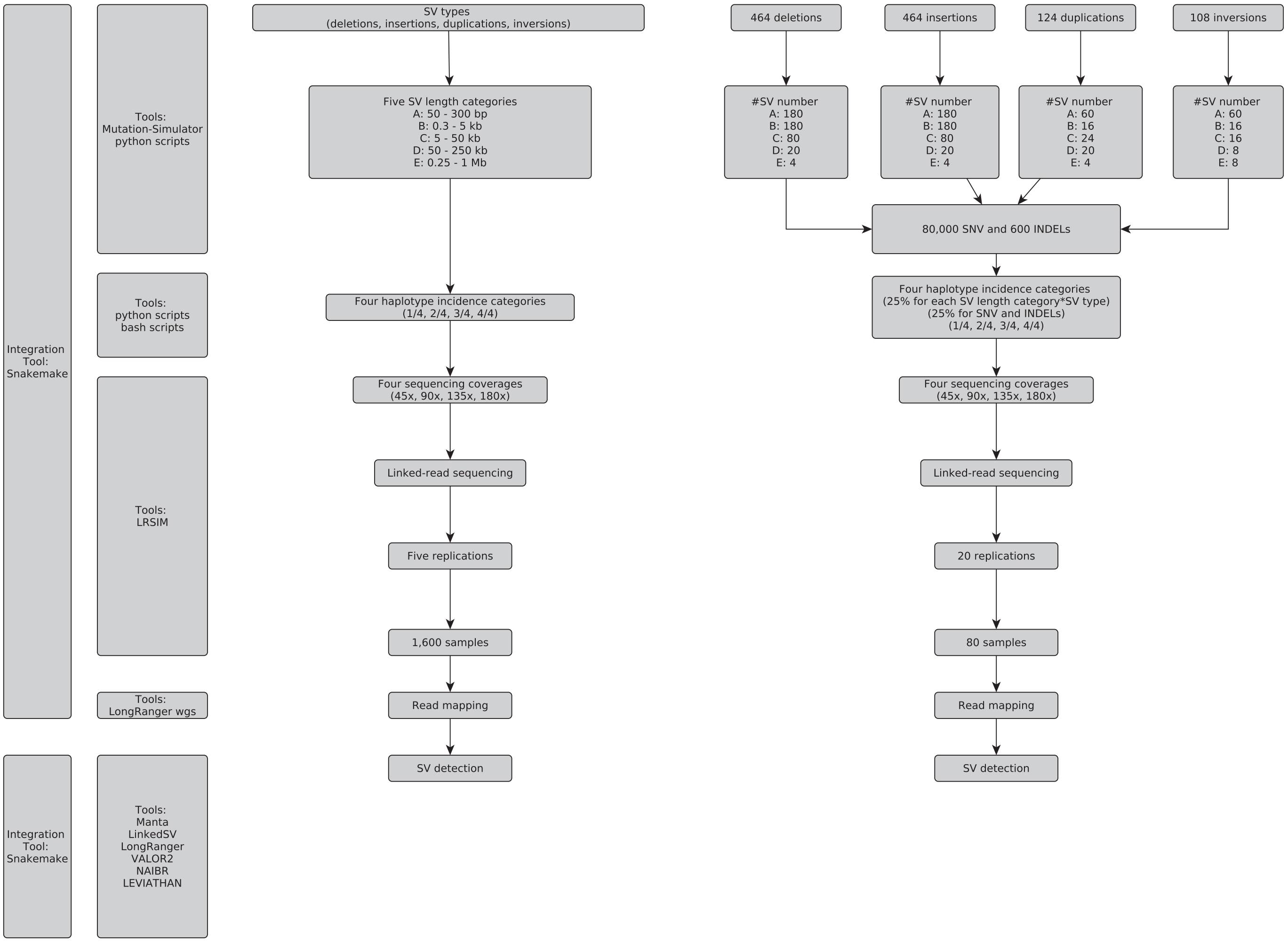
Integration Tool: Snakemake
Tools: Mutation-Simulator python scripts
Tools: python scripts bash scripts
Tools: LRSIM
Tools: LongRanger wgs
Integration Tool: Snakemake
Tools: Manta LinkedSV LongRanger VALOR2 NAIBR LEVIATHAN
SV types (deletions, insertions, duplications, inversions)
Five SV length categories A: 50 - 300 bp B: 0.3 - 5 kb C: 5 - 50 kb D: 50 - 250 kb E: 0.25 - 1 Mb
Four haplotype incidence categories (1/4, 2/4, 3/4, 4/4)
Four sequencing coverages (45x, 90x, 135x, 180x)
Linked-read sequencing
Five replications
1,600 samples
Read mapping
SV detection
464 deletions
464 insertions
124 duplications
108 inversions
#SV number A: 180 B: 180 C: 80 D: 20 E: 4
#SV number A: 180 B: 180 C: 80 D: 20 E: 4
#SV number A: 60 B: 16 C: 24 D: 20 E: 4
#SV number A: 60 B: 16 C: 16 D: 8 E: 8
80,000 SNV and 600 INDELs
Four haplotype incidence categories (25% for each SV length category*SV type) (25% for SNV and INDELs) (1/4, 2/4, 3/4, 4/4)
Four sequencing coverages (45x, 90x, 135x, 180x)
Linked-read sequencing
20 replications
80 samples
Read mapping
SV detection

Fig. 1. Overview of the workflow of this study including used bioinformatic tools (left) in the simple (center) and complex (right) simulations. Detailed information of the workflow can be found in the material and methods section and on https://github.com/mw-qggp/SV_simulation_potato.

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 (≥ 10 kb)	x (≥ 20 kb)
LongRanger	short + linked	x		x (≥ 30 kb)	x (≥ 30 kb)
VALOR2	linked	x (≥100 kb)		x (≥ 80 kb)	
NAIBR	linked	x		x	x
LEVIATHAN	linked	x (≥ 1 kb)		x (≥ 1 kb)	x (≥ 1 kb)
Novel-X	linked		x		

2.4. Evaluation of SV calling

We calculated the sensitivity (1), which is also called statistical power in other studies, precision (2), which corresponds to 1 - false discovery rate, and the F1-score (harmonic average of the precision and sensitivity) (3) as

$$Sensitivity = TP/(TP + FN) \qquad (1)$$

$$Precision = TP/(TP + FP) \qquad (2)$$

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

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 ≤ 50 bp. For the SV length category C, a TP SV had break points and length differences compared to the simulated SV of ≤ 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 ≤ 250 bp. For deletions and inversions of the SV length category D, ≤ 550 bp and ≤ 800 bp were chosen as threshold, respectively. For deletions and inversions of the SV length category E, ≤ 250 and ≤ 550 bp were used, respectively. For insertions, the start of a TP insertion had a break point that did differ ≤ 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) [42].

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. In contrast to the simple simulations, we only evaluated the performance of SV callers for the SV length categories C, D, and 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.

3. Results

Six linked-read and one short-read SV caller (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.

3.1. 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

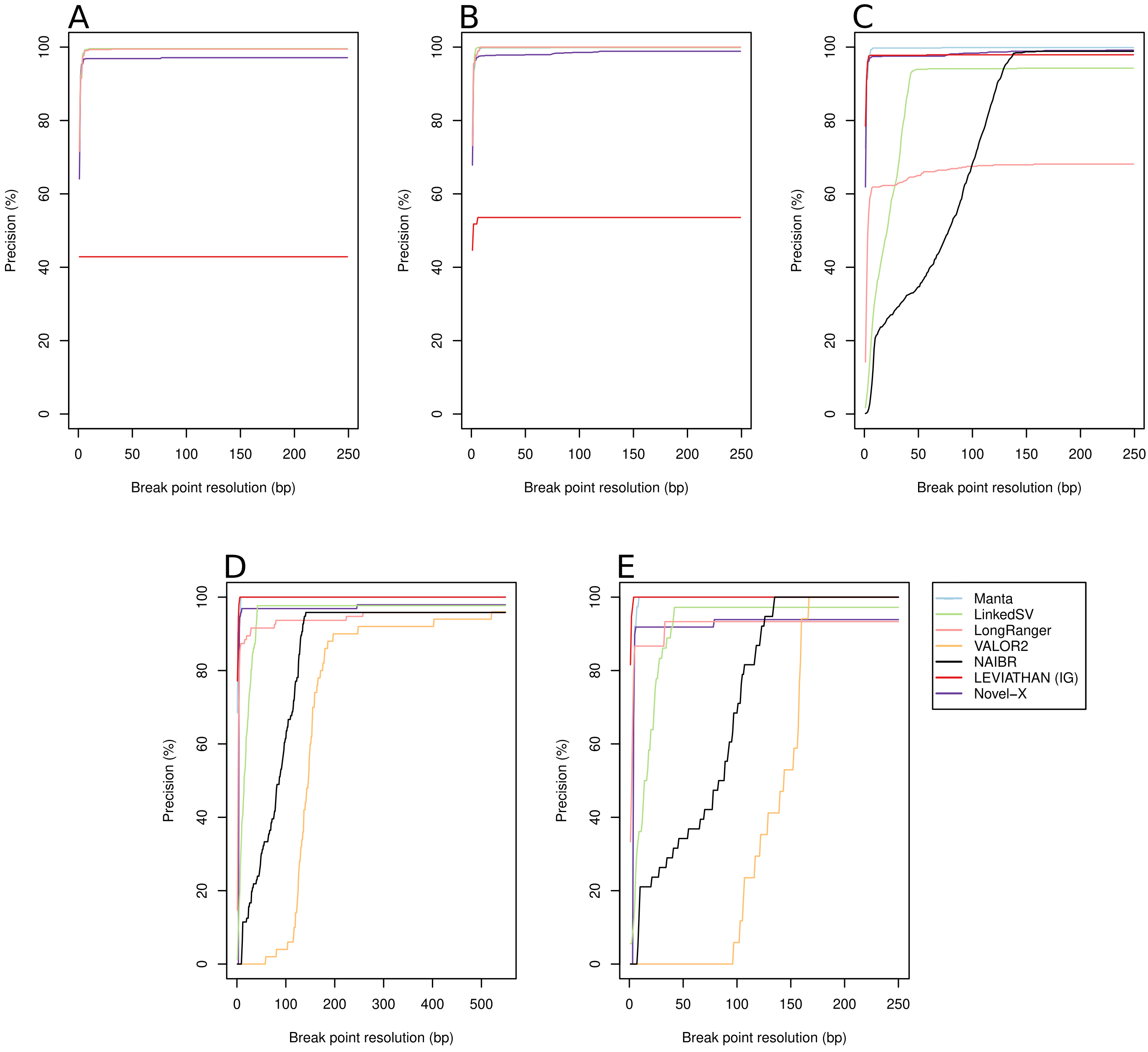

Fig. 2. Break point resolution in bp of the different SV callers for five structural variant (SV) length categories: A (50–300 bp), B (0.3–5 kb), C (5–50 kb), D (50–250 kb), E (0.25–1 Mb) based on the detection of homozygous (4/4) deletions (insertions, Novel-X) using a linked-read sequencing coverage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

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 ≤ 10 bp. In contrast, the BPR of LongRanger and VALOR2 were the lowest.

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

3.2. 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 (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% followed closely by LinkedSV (short) (95.9%, 95.6%, Fig. 3 III), 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 S6, S7). Larger deletions could be identified by linked-read SV callers (Supplementary Table S8 - S10). 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 category 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 as was observed for SV length category D, where Manta (89.6%) and NAIBR (88.5%) showed the highest average F1-scores.

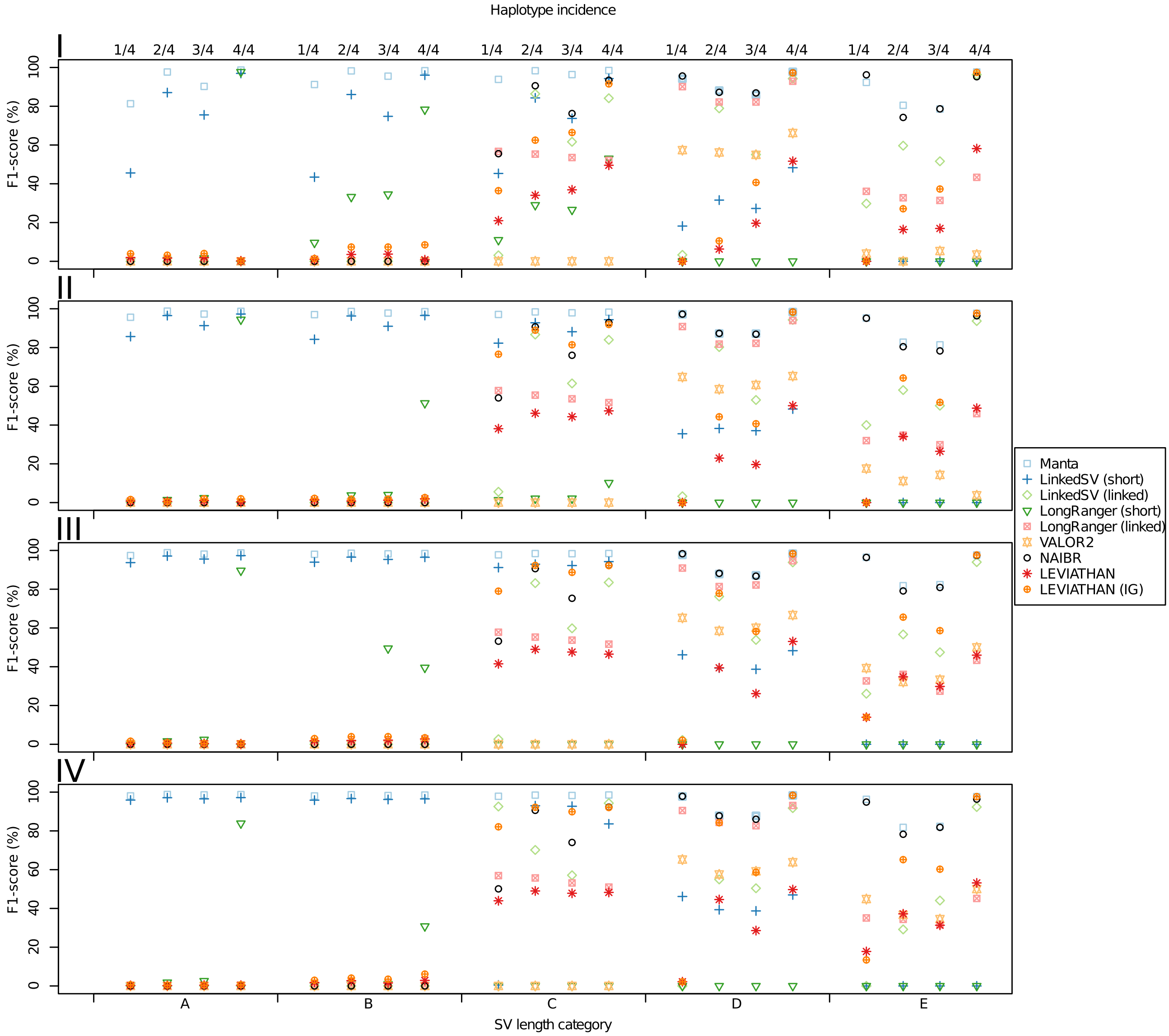

Fig. 3. F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of deletions of five structural variant (SV) length categories: A (50–300 bp), B (0.3–5 kb), C (5–50 kb), D (50–250 kb), E (0.25–1 Mb) and four haplotype incidences (1/4, 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.

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%) (Supplementary Fig. S3III, Supplementary Tables S11, S12), 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 S13). An even better performance of linked-read SV callers was observed for the SV length categories D and E (Supplementary Tables S14, S15), 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% (Supplementary Fig. S4III) 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 S16). 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

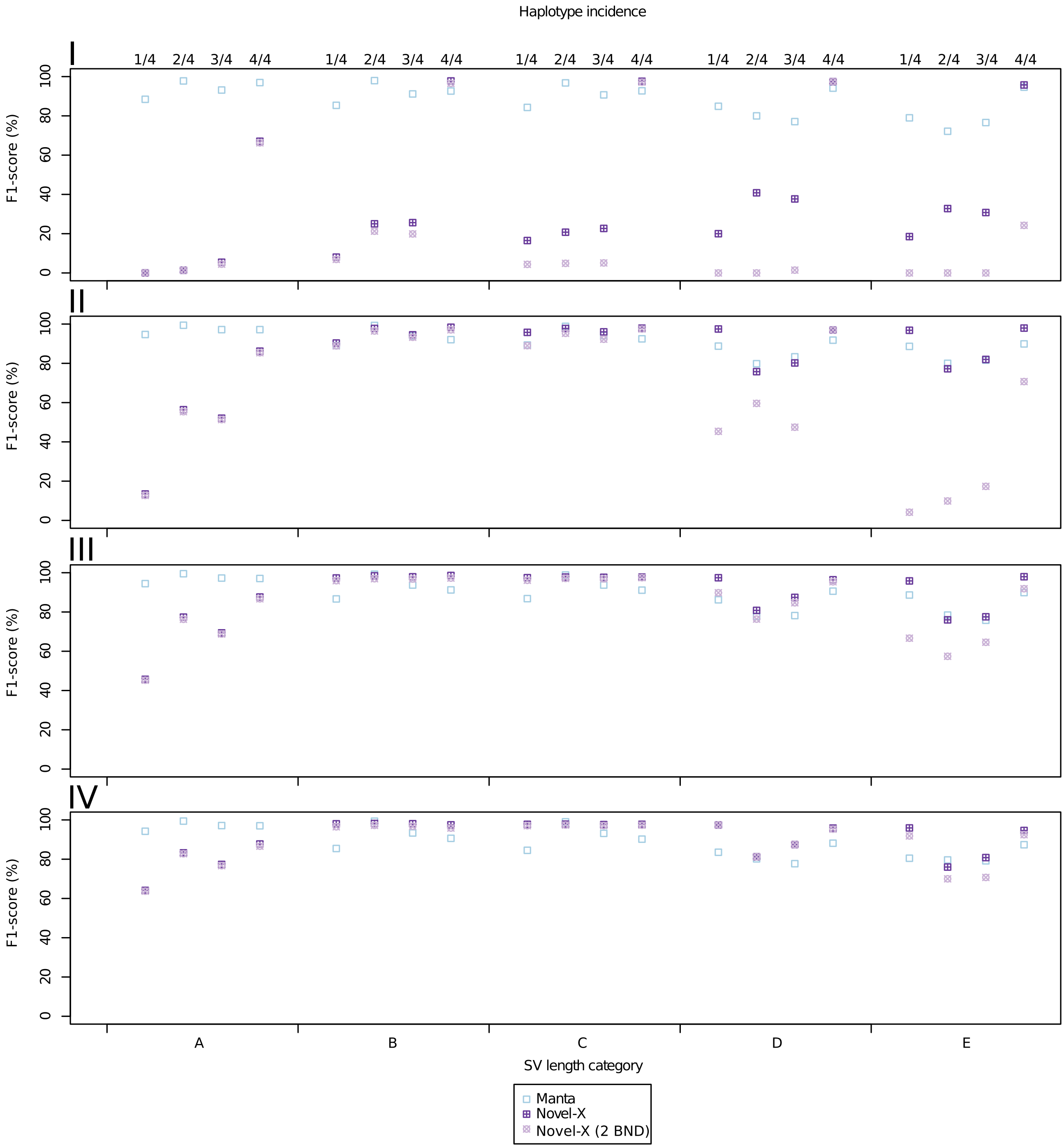

Fig. 4. F1-score, which is the harmonic mean of the precision and sensitivity, observed in the simple simulations, for the detection of insertions of five structural variant (SV) length categories: A (50–300 bp), B (0.3–5 kb), C (5–50 kb), D (50–250 kb), E (0.25–1 Mb) and four haplotype incidences (1/4, 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.

lower compared to those values observed for Manta (97.7%, 95.7%, 99.8%) (Supplementary Table S17). For the SV length category C, Manta performed well with an average F1-score of 97.2%. LEVIATHAN (IG) followed slightly behind with an average F1-score of 84.4%. LongRanger showed a considerably lower F1-score of 34.4% because of the low sensitivity (21.8%) (Supplementary Table S18). In contrast to the SV length category C, NAIBR and LinkedSV were able to detect duplications of the SV length category D (Supplementary Table S19). 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 S20), 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 the detection of 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. 4 III). 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 S21). 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 S22- S25).

3.3. 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 S6 - S10). 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 scenario 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. The only exception was the detection of insertions. The average F1-scores of Novel-X increased considerably with an increasing coverage.

3.4. SV detection assuming 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.

3.5. Uniquely detected SV by different SV callers

In addition to considering all simulated SV for the evaluations, we also performed evaluations of the SV that were uniquely detected by one SV caller. Manta showed a high number of uniquely detected SV compared to the linked-read SV callers (Fig. 5). Additionally, the total number of detected SV was also the highest for Manta compared to all other SV callers. The uniquely detected SV by Manta had high precisions between 95% and 100% for the different SV types (Fig. 6). In addition, high median values were also observed for LinkedSV (short) for deletions (87.5%) and for Novel-X for insertions (87.6%). The precisions of the uniquely detected SV for the other linked-read SV callers were with values below 20% considerably lower, but also their number was with values between one and 20 much lower compared to those of Manta.

3.6. Evaluation of SV detection using complex simulations

In addition to the simple simulations explained before, 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 (Supplementary Tables S2 - S5). 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. In contrast to the simple simulations, LongRanger (linked) showed lower sensitivity values for the detection of larger deletions.

4. Discussion

Due to tremendous improvements of sequencing technologies and bioinformatic tools, genome-wide SV detection became possible [19]. 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 [4,29], low precision and a lack of detecting large SV as well as assembling insertions were reported [5,19,22,35]. In contrast, SV calling based on long-read sequencing overcomes these issues but results in higher operational

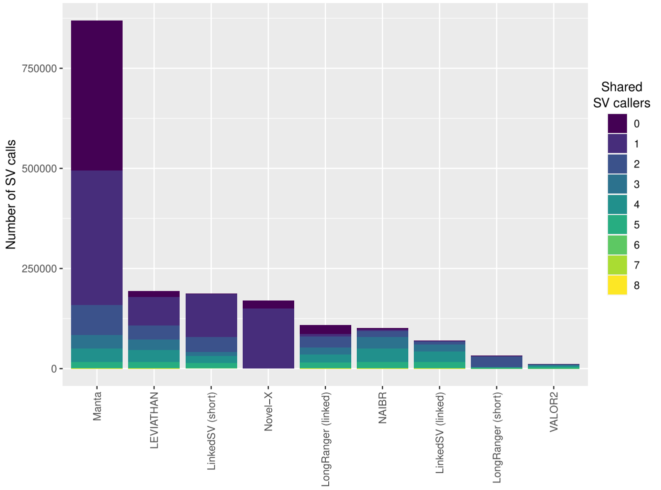

Fig. 5. Number of SV calls shared among SV callers where SV calls across all SV types, SV length categories, haplotype incidences, sequencing coverages, and repetitions were considered.

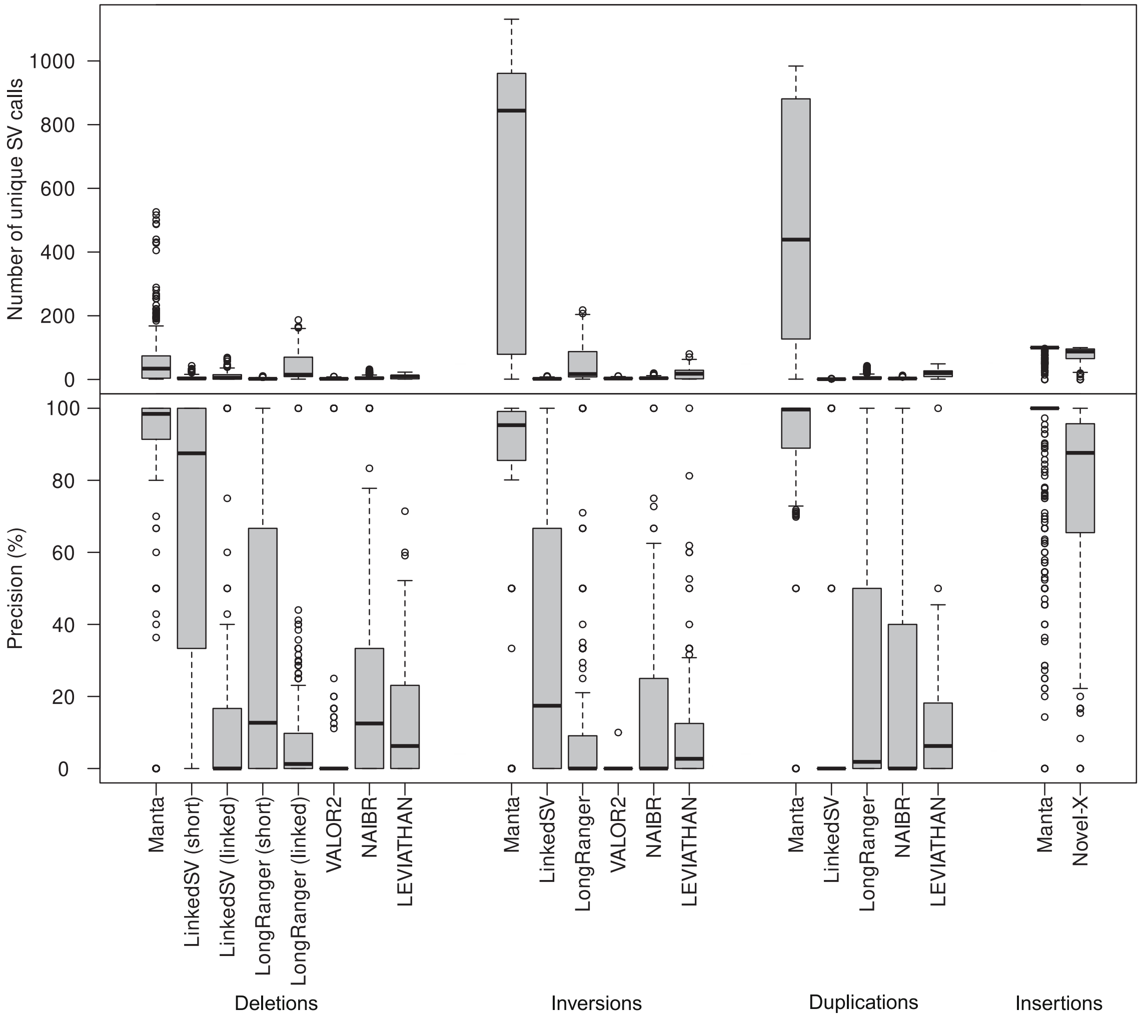

Fig. 6. Precision of uniquely detected SV by each SV caller. Only SV types and SV length categories of all scenarios (haplotype incidences, sequencing coverages, repetitions) for the simple simulations were considered which can be detected by the particular SV caller.

costs, large DNA input requirement, as well as lower sample throughput [19]. 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 [12]. Despite the discontinued support of 10xGenomics offering linked-read sequencing, many current studies are available where 10x linked-reads are used [45,49]. More importantly, linked-read sequencing is still offered by BGI as single tube long fragment reads (stLFR) [50]. Two previously described linked-read SV callers were not considered in our study, due to discontinued support and algorithm similarity to LongRanger (GROC-SVs) [46] or the functional restriction to human genomes (ZoomX) [55].

4.1. Simple vs. complex simulations

In general, the high sensitivity and precision values observed in the simple simulations were confirmed by the complex simulations. Therefore, only the results of the simple simulations were discussed in the following sections. Furthermore, in both, simple and complex simulations, 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 occurring simultaneously have not a negative influence on the detection of each other. Therefore, the high precision values observed in our complex simulations can be also expected in experimental data of tetraploid potato varieties.

4.2. 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. In contrast to using linked-read SV callers, linked-read signals are, except for the mapping of the reads, simply not considered by the short-read SV callers to call SV. Therefore, we used Manta as short-read SV caller to evaluate the detection of SV using short-read signals based on linked-read mapping.

We observed high precision and sensitivity values for the SV detection using the short-read SV callers Manta and LinkedSV (short) (Fig. 3, Supplementary Tables S6 - S10). Our observations are supported by recent comprehensive SV calling evaluation studies in humans [4,29]. 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 [13,44,45,48]. One reason might be that the latter studies evaluated SV callers that have been developed ten years ago such as Pindel [7] or BreakDancer [1]. The latter 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 [52]. 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 [34].

In our study, the F1-score of the short-read SV caller Manta was always equal or higher compared to the linked-read SV callers NAIBR or LinkedSV, caused by a lower sensitivity of the linked-read SV callers. In contrast, the precision was high for short- and linked-read SV callers. However, only Manta, LinkedSV (short), and Novel-X showed high precision values for uniquely detected SV (Figs. 5, 6). The low precision values of uniquely detected SV for linked-read SV callers is due to the low number of uniquely detected SV (Fig. 6) by those. In contrast, the high precision of linked-read SV callers considering all simulated SV can be explained by the usage of short-read signals and barcode information which was also previously reported in human [45]. 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 [45]. In contrast, Fang et al. [14] compared the performance of linked-read SV callers to the short-read SV callers Lumpy [31] and Delly [41] and showed that the F1-score was higher for NAIBR and LinkedSV than for Delly and Lumpy. This observation can be explained thereby that Manta showed a better performance to detect SV in human [4,29] and barley [52] compared to Delly and Lumpy. However, 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.

4.3. 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, a 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, Supplementary Fig. S3, S4). 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 examined linked-read SV callers were developed for the detection of large SV ($\geq$ 10 kb) [14,58] and the focus laid not on the detection of short SV. However, NAIBR and LEVIATHAN were able to detect SV between 1 and 5 kb in the human genome, but they showed a low sensitivity [12,37]. This finding is in agreement with 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 overlapped 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 examined 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 S6 - S25). 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 for the detection of large SV was reported [14]. The high precision values to detect large deletions and inversions in the human genome reported for VALOR2 [24] could be supported by our results as well (Supplementary Tables S9, S10, S14, S15). 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).

4.4. 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 data. The strongest differences were observed for calling deletions of the SV length category A (Fig. 3, Supplementary Table S6) from 45x (~11x per haplotype in potato) to 90x (~22x per haplotype) sequencing coverage, 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. [4] who reported a higher sensitivity for short-read SV callers using higher levels of sequencing coverage. In detail, these authors reported above 30x (15x per haplotype) that the sensitivity increased marginally whereas below 30x the sensitivity decreased considerably. 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 [29]. 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 seems to be marginal (Figs. 3, 4, Supplementary Fig. S3, S4). 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 given by the molecule signals. In contrast, for short-read sequencing, only reads can be considered where the sequence covered the break points. Therefore, the reduction of the sequencing coverage results in fewer short-read signals which has more severe consequences for the SV detection compared to linked-read signals.

In contrast to the above described trend, we have observed two exceptions where the sequencing coverage influenced the SV detection for linked-read SV callers. First, detecting insertions by Novel-X is strongly influenced by the sequencing coverage (Fig. 4). 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 [35]. 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 S8) or inversions (20.4%) (Supplementary Table S13). 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 [37]. However, as explained above, short-read signals benefit from an increased coverage.

4.5. 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 HIs 1/4 and 4/4 (Figs. 3, 4, Supplementary Fig. S3, S4). The reason for this observation remains elusive and additional research is needed in the field of polyploid SV calling.

Approaches for SV genotyping based on short-read sequencing have been described for diploid genomes [18] even though it is more complex [3] compared to well established SNV genotyping based on read depth signals [40]. 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 [10]. Considering the need of improvements of diploid SV genotyping [6,27] and the issues of polyploid SNV genotyping [10], polyploid SV genotyping will be one of the big challenges in crop research.

4.6. 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) [20,25,26,43] caused by the limited anchor size due to the short insert size of the sequencing library and the corresponding incapacity to span over larger repetitive regions in the genome [35]. Manta is able to determine the SV length for insertions up to ~1 kb. However, SV calling based on linked-read sequencing can principally detect larger insertions. But, up to date, only one algorithm (Novel-X) was developed for the detection of insertions.

As this algorithm revealed high sensitivity and precision values to detect insertions in our study (Fig. 4, Supplementary Tables S21 - S25), 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. [35] 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.

4.7. 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 for two potato chromosomes (Table 2). 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.

5. Conclusion

We observed high precision and sensitivity values considering different sequencing coverages for the SV detection in the potato genome. Our observations highlighted the importance of short-read signals by Manta and LinkedSV to detect short SV, whereas Manta and NAIBR performed well for detecting larger deletions, inversions, and duplications. 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 scenario, the highest precision and sensitivity values were observed. Finally, the short-read algorithms were stronger influenced by the sequencing coverage than the linked-read SV callers, except Novel-X, where at least a sequencing coverage of about 22x per haplotype should be used to detect insertions.

Table 2
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.

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.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

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. We thank Julien Bruckmüller, Rafal Zgadzaj, and Stefanie Hartje for comments and suggestions on an earlier version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2023.110568.

References

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

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

[3] D.L. Cameron, J. Schröder, J.S. Penington, H. Do, R. Molania, A. Dobrovic, T. P. Speed, A.T. Papenfuss, GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly, Genome Res. 27 (2017) 1–11.

[4] D.L. Cameron, L. Di Stefano, A.T. Papenfuss, Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software, Nat. Commun. 10 (2019) 3240.

[5] M.J. Chaisson, J. Huddleston, M.Y. Dennis, P.H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J.M. Landolin, J. A. Stamatoyannopoulos, M.W. Hunkapiller, J. Korlach, E.E. Eichler, Resolving the

complexity of the human genome using single-molecule sequencing, Nature 517 (2015) 608–611.

[6] V. Chander, R.A. Gibbs, F.J. Sedlazeck, Evaluation of computational genotyping of structural variation for clinical diagnoses, GigaScience 8 (2019) 1–7.

[7] K. Chen, J.W. Wallis, M.D. McLellan, D.E. Larson, J.M. Kalicki, C.S. Pohl, S. D. McGrath, M.C. Wendl, Q. Zhang, D.P. Locke, X. Shi, R.S. Fulton, T.J. Ley, R. K. Wilson, L. Ding, E.R. Mardis, BreakDancer: an algorithm for high-resolution mapping of genomic structural variation, Nat. Methods 6 (2009) 677–681.

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

[9] C. Chiang, A.J. Scott, J.R. Davis, E.K. Tsang, X. Li, Y. Kim, T. Hadzic, F.N. Damani, L. Ganel, S.B. Montgomery, A. Battle, D.F. Conrad, I.M. Hall, The impact of structural variation on human gene expression, Nat. Genet. 49 (2017) 692–699.

[10] D.P. Cooke, D.C. Wedge, G. Lunter, Benchmarking small-variant genotyping in polyploids, Genome Res. 32 (2022) 403–408.

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

[12] R. Elyanow, H.T. Wu, B.J. Raphael, Identifying structural variants using linked-read sequencing data, Bioinformatics 34 (2018) 353–360.

[13] A.C. English, W.J. Salerno, O.A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D. I. Ritter, C.R. Beck, C.F. Davis, M. Dahdouli, S. Ma, A. Carroll, N. Veeraraghavan, J. Bruestle, B. Drees, A. Hastie, E.T. Lam, S. White, P. Mishra, M. Wang, Y. Han, F. Zhang, P. Stankiewicz, D.A. Wheeler, J.G. Reid, D.M. Muzny, J. Rogers, A. Sabo, K.C. Worley, J.R. Lupski, E. Boerwinkle, R.A. Gibbs, Assessing structural variation in a personal genome-towards a human reference diploid genome, BMC Genomics 16 (2015) 286.

[14] L. Fang, C. Kao, M.V. Gonzalez, F.A. Mafra, R. Pellegrino da Silva, M. Li, S. S. Wenzel, K. Wimmer, H. Hakonarson, K. Wang, LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data, Nat. Commun. 10 (2019) 5585.

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

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

[17] M. Göktay, A. Fulgione, A.M. Hancock, 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, Mol. Biol. Evol. 38 (2020) 1498–1511.

[18] G. Hickey, D. Heller, J. Monlong, J.A. Sibbesen, J. Sirén, J. Eizenga, E.T. Dawson, E. Garrison, A.M. Novak, B. Paten, Genotyping structural variants in pangenome graphs using the vg toolkit, Genome Biol. 21 (2020) 35.

[19] S.S. Ho, A.E. Urban, R.E. Mills, Structural variation in the sequencing era, Nat. Rev. Genet. 21 (2020) 171–189.

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

[21] Y. Hu, V. Colantonio, B.S. Müller, K.A. Leach, A. Nanni, C. Finegan, B. Wang, M. Baseggio, C.J. Newton, E.M. Juhl, L. Hislop, J.M. Gonzalez, E.F. Rios, L. C. Hannah, K. Swarts, M.A. Gore, T.A. Hennen-Bierwagen, A.M. Myers, A. M. Settles, W.F. Tracy, M.F. Resende, Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn, Nat. Commun. 12 (2021) 1227.

[22] J. Huddleston, E.E. Eichler, An incomplete understanding of human genetic variation, Genetics 202 (2016) 1251–1254.

[23] M. Iovene, T. Zhang, Q. Lou, C.R. Buell, J. Jiang, Copy number variation in potato - an asexually propagated autotetraploid species, Plant J. 75 (2013) 80–89.

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

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

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

[27] M.M. Khayat, S. Mohammad, E. Sahraeian, S. Zarate, A. Carroll, H. Hong, B. Pan, L. Shi, R.A. Gibbs, M. Mohiyuddin, Y. Zheng, F.J. Sedlazeck, Hidden biases in germline structural variant detection, Genome Biol. 22 (2021) 347.

[28] J. Köster, F. Mölder, K.P. Jablonski, B. Letcher, M.B. Hall, C.H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S.O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, Sustainable data analysis with Snakemake, F1000Research 10 (2021) 33.

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

[30] M.A. Kühl, B. Stich, D.C. Ries, Mutation-simulator: fine-grained simulation of random mutations in any genome, Bioinformatics 37 (2021) 568–569.

[31] R.M. Layer, C. Chiang, A.R. Quinlan, I.M. Hall, LUMPY: a probabilistic framework for structural variant discovery, Genome Biol. 15 (2014) R84.

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

[33] R. Luo, F.J. Sedlazeck, C.A. Darby, S.M. Kelly, M.C. Schatz, LRSim: a linked-reads simulator generating insights for better genome partitioning, Comp. Struct. Biotechnol. J. 15 (2017) 478–484.

[34] P. Marks, S. Garcia, A.M. Barrio, K. Belhocine, J. Bernate, R. Bharadwaj, K. Bjornson, C. Catalanotti, J. Delaney, A. Fehr, I.T. Fiddes, B. Galvin, H. Heaton, J. Herschleb, C. Hindson, E. Holt, C.B. Jabara, S. Jett, N. Keivanfar, S. Kyriazopoulou-Panagiotopoulou, M. Lek, B. Lin, A. Lowe, S. Mahamdallie, S. Maheshwari, T. Makarewicz, J. Marshall, F. Meschi, C.J. O'Keefe, H. Ordonez, P. Patel, A. Price, A. Royall, E. Ruark, S. Seal, M. Schnall-Levin, P. Shah, D. Stafford, S. Williams, I. Wu, A.W. Xu, N. Rahman, D. MacArthur, D.M. Church, Resolving the full spectrum of human genome variation using linked-reads, Genome Res. 29 (2019) 635–645.

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

[36] S.G. Milner, M. Jost, S. Taketa, E.R. Mazón, A. Himmelbach, M. Oppermann, S. Weise, H. Knüpffer, M. Basterrechea, P. König, D. Schüler, R. Sharma, R. K. Pasam, T. Rutten, G. Guo, D. Xu, J. Zhang, G. Herren, T. Müller, S.G. Krattinger, B. Keller, Y. Jiang, M.Y. González, Y. Zhao, A. Habekuß, S. Färber, F. Ordon, M. Lange, A. Börner, A. Graner, J.C. Reif, U. Scholz, M. Mascher, N. Stein, Genebank genomics highlights the diversity of a global barley collection, Nat. Genet. 51 (2019) 319–326.

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

[38] P. Morisse, C. Lemaitre, F. Legeai, LRez: C ++ API and toolkit for analyzing and managing linked-reads data, Bioinformatics Advances 1 (2021) vbab022.

[39] H. Nishida, T. Yoshida, K. Kawakami, M. Fujita, B. Long, Y. Akashi, D.A. Laurie, K. Kato, 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 (2013) 27–37.

[40] R. Poplin, V. Ruano-Rubio, M.A. DePristo, T.J. Fennell, M.O. Carneiro, G.A. Van der Auwera, D.E. Kling, L.D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M.J. Daly, B. Neale, D. G. MacArthur, E. Banks, Scaling accurate genetic variant discovery to tens of thousands of samples, bioRxiv (2017) 201178.

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

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

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

[44] F.J. Sedlazeck, H. Lee, C.A. Darby, M.C. Schatz, Piercing the dark matter: bioinformatics of long-range sequencing and mapping, Nat. Rev. Genet. 19 (2018) 329–346.

[45] R. Sethi, J. Becker, J. de Graaf, M. Löwer, M. Suchan, U. Sahin, D. Weber, Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions, PLoS Comput. Biol. 16 (2020), e1008397.

[46] N. Spies, Z. Weng, A. Bishara, J. McDaniel, D. Catoe, J.M. Zook, M. Salit, R.B. West, S. Batzoglou, A. Sidow, Genome-wide reconstruction of complex structural variants using read clouds, Nat. Methods 14 (2017) 915–920.

[47] B. Stich, D.V. Inghelandt, Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato, Front. Plant Sci. 9 (2018) 159.

[48] P.H. Sudmant, T. Rausch, E.J. Gardner, R.E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M.H.Y. Fritz, M.K. Konkel, A. Malhotra, A.M. Stütz, X. Shi, F.P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M.J. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H.Y. Lam, X.J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J.M. Kidd, Y. Kong, E.W. Lameijer, S. McCarthy, P. Flicek, R.A. Gibbs, G. Marth, C.E. Mason, A. Menelaou, D. M. Muzny, B.J. Nelson, A. Noor, N.F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E.E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A.A. Shabalin, A. Untergasser, J.A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M.A. Batzer, S.A. McCarroll, R.E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S.E. Devine, C. Lee, E.E. Eichler, J.O. Korbel, An integrated map of structural variation in 2,504 human genomes, Nature 526 (2015) 75–81.

[49] K. Talsania, T.W. Shen, X. Chen, E. Jaeger, Z. Li, Z. Chen, W. Chen, B. Tran, R. Kusko, L. Wang, A.W.C. Pang, Z. Yang, S. Choudhari, M. Colgan, L.T. Fang, A. Carroll, J. Shetty, Y. Kriga, O. German, T. Smirnova, T. Liu, J. Li, B. Kellman, K. Hong, A.R. Hastie, A. Natarajan, A. Moshrefi, A. Granat, T. Truong, R. Bombardi, V. Mankinen, D. Meerzaman, C.E. Mason, J. Collins, E. Stahlberg, C. Xiao, C. Wang, W. Xiao, Y. Zhao, Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies, Genome Biol. 23 (2022) 255.

[50] O. Wang, R. Chin, X. Cheng, M.K. Yan Wu, Q. Mao, J. Tang, Y. Sun, E. Anderson, H. K. Lam, D. Chen, Y. Zhou, L. Wang, F. Fan, Y. Zou, Y. Xie, R.Y. Zhang, S. Drmanac, D. Nguyen, C. Xu, C. Villarosa, S. Gablenz, N. Barua, S. Nguyen, W. Tian, J.S. Liu, J. Wang, X. Liu, X. Qi, A. Chen, H. Wang, Y. Dong, W. Zhang, A. Alexeev, H. Yang, J. Wang, K. Kristiansen, X. Xu, R. Drmanac, B.A. Peters, Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules

enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly, Genome Res. 29 (2019) 798–808.

[51] N.I. Weisenfeld, V. Kumar, P. Shah, D.M. Church, D.B. Jaffe, Direct determination of diploid genome sequences, Genome Res. 27 (2017) 757–767.

[52] M. Weisweiler, C. Arlt, P.Y. Wu, D.V. Inghelandt, T. Hartwig, B. Stich, Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation, Theor. Appl. Genet. 135 (2022) 3511–3529.

[53] A.M. Wenger, P. Peluso, W.J. Rowell, P.C. Chang, R.J. Hall, G.T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N.D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.S. Chin, A.M. Phillippy, M.C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F.J. Sedlazeck, J.M. Zook, H. Li, S. Koren, A. Carroll, D.R. Rank, M.W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome, Nat. Biotechnol. 37 (2019) 1155–1162.

[54] P.Y. Wu, B. Stich, S. Hartje, K. Muders, V. Prigge, D. Van Inghelandt, Optimal implementation of genomic selection in clone breeding programs - exemplary in potato: I. Effect of breeding strategy, implementation stage, and selection intensity on genetic gain, bioRxiv (2022), https://doi.org/10.1101/2022.11.25.517496.

[55] L.C. Xia, J.M. Bell, C. Wood-Bouwens, J.J. Chen, N.R. Zhang, H.P. Ji, Identification of large rearrangements in cancer genomes with barcode linked reads, Nucleic Acids Res. 46 (2018), e19.

[56] X. Xu, X. Liu, S. Ge, J.D. Jensen, F. Hu, X. Li, Y. Dong, R.N. Gutenkunst, L. Fang, L. Huang, J. Li, W. He, G. Zhang, X. Zheng, F. Zhang, Y. Li, C. Yu, K. Kristiansen, X. Zhang, J. Wang, M. Wright, S. McCouch, R. Nielsen, J. Wang, W. Wang, Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes, Nat. Biotechnol. 301 (2012) 105–111.

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

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