bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Comparison of read mapping and variant calling tools for the analysis of plant NGS data

Hanna Marie Schilbert[1], Andreas Rempel[1,2], Boas Pucker[1,3]

[1]Genetics and Genomics of Plants, CeBiTec and Faculty of Biology, Bielefeld University, Bielefeld, Germany

[2]Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Faculty of Technology, Bielefeld University, Bielefeld, Germany

[3]Molecular Genetics and Physiology of Plants, Faculty of Biology and Biotechnology, Ruhr-University Bochum, Bochum, Germany

hschilbe@cebitec.uni-bielefeld.de

https://orcid.org/0000-0003-0474-7753

arempel@cebitec.uni-bielefeld.de

https://orcid.org/0000-0003-0609-5621

bpucker@cebitec.uni-bielefeld.de

https://orcid.org/0000-0002-3321-7471

Corresponding author: Hanna Marie Schilbert, hschilbe@cebitec.uni-bielefeld.de

Key words: single nucleotide variant, SNV, SNP, insertion/deletion, InDel, population genomics, re-sequencing

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

High-throughput sequencing technologies have rapidly developed during the past years and became an essential tool in plant sciences. However, the analysis of genomic data remains challenging and relies mostly on the performance of automatic pipelines. Frequently applied pipelines involve the alignment of sequence reads against a reference sequence and the identification of sequence variants. Since most benchmarking studies of bioinformatics tools for this purpose have been conducted on human datasets, there is a lack of benchmarking studies in plant sciences.

In this study we evaluated the performance of 50 different variant calling pipelines, including five read mappers and ten variant callers, on six real plant datasets of the model organism *Arabidopsis thaliana*. Sets of variants were evaluated based on various parameters including sensitivity and specificity. We found that all investigated tools are suitable for analysis of NGS data in plant research. When looking at different performance metrices, BWA-MEM and Novoalign were the best mappers and GATK returned the best results in the variant calling step.

Introduction

As the basis of biological properties and heredity, the genome of a species is a valuable resource for numerous studies. However, there are subtle differences between individuals of the same species, which are of academic and economic interest as these determine phenotypic differences. Dropping sequencing costs boosted high-throughput sequencing projects, thus facilitating the analysis of this genetic diversity. Variations within the *A. thaliana* population were studied in the 1001 genomes project[1]. As the number of high-quality reference genome sequences rises continuously, the number of re-sequencing projects increases as well[2]. There are pan-genome projects for various species focusing on the genome evolution[3–5] and mapping-by-sequencing projects which focus on agronomically important traits of crops[6–9].

An accurate and comprehensive identification of sequence variants between a sample and the reference sequence is the major challenge in many re-sequencing projects[10]. The large amount and diverse nature of NGS-data types (as reviewed in[11]), the diversity of bioinformatics algorithms, and the quality of the reference genome sequence render the choice of the best approach challenging.

Variant calling pipelines often start with (1) the preprocessing of sequence reads, followed by (2) the alignment (mapping) of these reads to a reference sequence. Finally, the (3) identification (calling) of sequence variants is performed based on alignments. Each of these three steps can be carried out by various alternative programs using different algorithms, which influence the accuracy and sensitivity of the resulting variant set.

First, read processing can be required if the read quality is at least partially low. Some downstream tools require that sequence reads come with quality scores in a certain system, namely phred33 or phred64. The conversion between different systems is allowed by some read processing tools. Popular read processing tools are FastQC[12], htSeqTools[13], NGSQC[14], SAMStat[15], and Trimmomatic[16].

As the read mapping determines the quality of the alignment, it is arguably the most important step[10]. Sequence reads are aligned to a suitable, but not necessarily the best place in the genome sequence. Often, there is a trade-off between mapping speed and the quality of the resulting alignment[17].

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Numerous mappers are available, which utilize different algorithms and criteria to generate alignments[18,10]. Consequently, the choice of tool and parameters can have a large influence on the outcome of the mapping[19,20]. Reads originating from PCR duplicates should be remove from the mapping prior to the variant calling to improve reliability of the results[20]. Moreover, the quality of the reference genome sequence plays an important role for the performance of the mapper. Particular challenges are low-complexity sequences, repetitive regions, collapsed copies of sequences, contaminations, or gaps in the reference genome sequence[10]. Frequently applied read mapper are Bowtie2[21], BWA-MEM[22], CLC Genomics Workbench (Quiagen), GEM3[23], Novoalign (http://novocraft.com/), and SOAP2[24]. While most of these tools are freely available for academic use as command line versions, CLC Genomics Workbench is a proprietary software suit for genomics with a graphical user interface. Detailed characteristics and algorithms of each mapper have been described elsewhere[25,26,20,18].

Finally, genomic variants like single nucleotide variants (SNVs) or small insertions/deletions (InDels) can be inferred by variant callers based on sequence read alignment. Popular variant callers like SAMtools/BCFtools[27], CLC Genomics Workbench (Quiagen), FreeBayes[28], GATK[29–32], LoFreq[33], SNVer[34], VarDict[35], and VarScan[36] use a variety of different approaches to call variants. Consequently, resulting variant sets differ depending on the employed methods (e.g. Bayesian), which come with strengths and weaknesses concerning the identification of specific variant types ([37], Table 19.5; [10], Table 4). Several factors that contribute to high accuracy of variant callings are: (I) a high coverage of the variant position resulting in support for SNVs by several overlapping reads[38], (II) a careful study design[20], (III) joint variant calling for multiple samples to allow mutual support of genotypes[39].

The initial set of putative sequence variants is usually filtered to remove unreliable variant calls. Possible reasons for the identification of these variants in the first place are incorrect alignments, sequencing errors, or low-quality scores[10]. Read depth, mapping quality, and bias in the alignment to both strands are central criteria used in the filtering step. While this filter step aims to reduce the number of false positives, it simultaneously increases the number of false negatives. The best trade-off between sensitivity and specificity depends on the purpose of the respective study.

Many underlying algorithms of variant calling pipelines were developed for the analysis of variants in the human genome, e.g. to investigate genetic disorders or to study tumor samples[20,40–43]. Although the applications in biomedical research and plant sciences differ substantially, plant scientists have largely followed benchmarking studies derived from research on human samples assuming similar performances. Therefore, no comprehensive benchmarking study of read mapping and variant calling tools for plant genome sequences is described in the literature. Due to substantial differences in the nucleotide composition, a dedicated benchmarking on plant genomes sequences is advised. A recent study compared the performance of BWA-MEM[22], SOAP2[24], and Bowtie2[21] with the two variant callers GATK[29–32] and SAMtools/BCFtools[27] on simulated and real tomato datasets[44]. To expand the sparse knowledge about the performance of other read mapping and variant calling tools on plant data, we set out to perform a systematic comparison. Our study evaluated the performance of 50 variant calling pipelines, testing combinations of five read mappers (Bowtie2, BWA-MEM, CLC-mapper, GEM3,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Novoalign) and eight different variant callers (SAMtools/BCFtools, CLC-caller, FreeBayes, GATK v3.8/v4.0/v4.1, LoFreq, SNVer, VarDict, VarScan) that are frequently applied in re-sequencing studies. Many combinations perform almost equally well on numerous data sets of the plant model organism *A. thaliana*. Illumina sequence reads were used for the detection of variants and provide the foundation for the comparison of these pipelines. Independent PacBio long reads were harnessed for the validation of identified variants based on an orthogonal sequencing technology.

Material and Methods

Sequence read data sets

We used paired-end Illumina reads from two different *A. thaliana* accessions, namely Columbia-0 (Col-0) and Niederzenz-1 (Nd-1) (Supplementary Table 1). The read quality was checked via FastQC[12] (Supplementary Table 1). Trimmomatic[16] was applied for light trimming and quality conversion to a Phred score of 33 if applicable. These data sets cover different Illumina sequencing platforms including GAIIX, MiSeq, and HiSeq 1500. While two data sets are the paired-end proportions of mate pair sequencing libraries (SRR2919288 and SRR3340911), these samples are 2×250 nt paired-end libraries.

Sequence read mapping

We chose five popular read mappers, namely Bowtie2 v2.3.4.3[21], BWA-MEM v0.7.17[22], CLC v11 Genomics Workbench (Quiagen), GEM3 v3.6[23], and Novoalign v3.09.01 (http://novocraft.com/) for this analysis. While most of these mappers are freely available for academic use, CLC is a proprietary software suit for genomic analyses. Paired-end reads were mapped against the TAIR10 reference genome sequence[45]. The executed commands for each tool can be found in Supplementary Table 2. Samtools v.1.8[27] was deployed for sorting of the BAM files. Reads originating from PCR duplicates where removed via MarkDuplicates in Picard-bf40841 (https://broadinstitute.github.io/picard/). Read groups or InDel qualities were assigned as these are required by some tools for downstream processing. While the plastome and chondrome sequences were included during the mapping step, variant caller performance was only assessed for the five nucleome sequences.

Variant calling

All mapping results were subjected to variant calling via CLC v11 Genomics Workbench (Quiagen), FreeBayes v1.2.0[28], Genome Analysis Tool Kit v3.8/v4.0/v4.1 HaplotypeCaller (GATK-HC)[29–32], LoFreq v2.1.3.1[33], Samtools v1.9[27] in combination with BCFtools v1.9 (alias BCFtools in the following), SNVer[34], VarDict[35], and VarScan[36]. We evaluated three different versions of GATK in order to analyze whether the applied version has a high impact on the variant calling pipeline performance. The executed commands for each tool can be found in Supplementary Table 2.

Performance measure of variant calling pipelines

The overall workflow of our benchmarking study is presented in Figure 1. We applied a previously described pipeline to validate sequence variants against the Nd-1 *de novo* assembly based on PacBio

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reads (https://github.com/bpucker/variant_calling)[46], which is crucial in order to assess the performance of each variant calling pipeline. A gold standard was generated from all validated variants by combining them into a single VCF file (https://docs.cebitec.uni-bielefeld.de/s/GG4CYJ7PcEwMFAF). Afterwards, the numbers of true positives, true negatives, false positives, and false negatives were calculated based on the gold standard and the initial VCF files for each variant calling pipeline and data set. Next, performance statistics including sensitivity, specificity, precision, accuracy, and F1 score were calculated per combination of mapper, variant caller, and data set (Table 1).

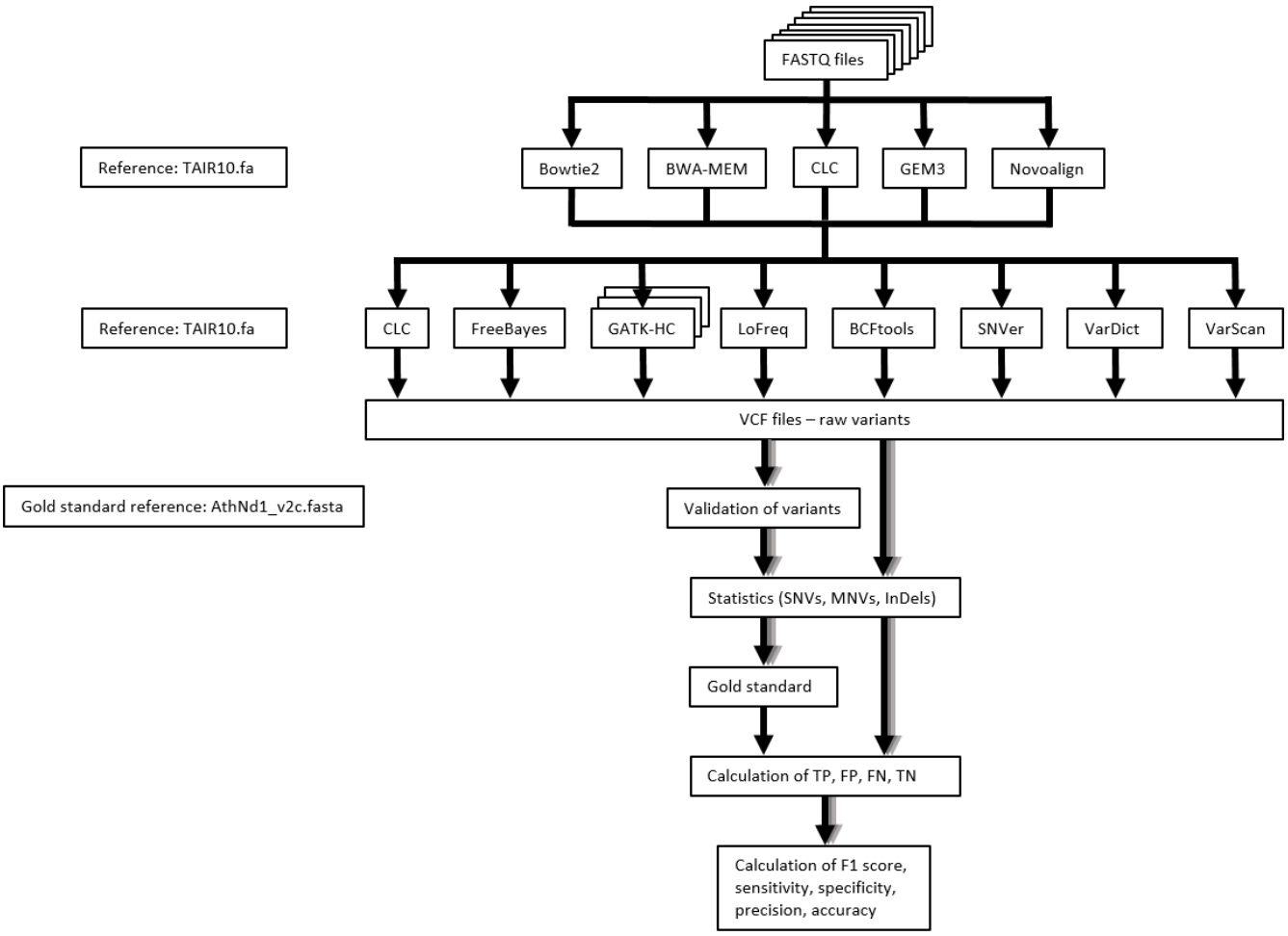

Figure 1: Workflow for the performance analysis of variant calling pipelines. First, reads within supplied FASTQ files were mapped against the TAIR10 A. thaliana reference genome sequence. Next, variants were called and saved in VCF files. All variants were subjected to a previously described validation process based on the Nd-1 genome sequence[46]. A gold standard was generated based on all validated variants. The initial variants called by each combination of mapper and variant caller were evaluated by analyzing whether they are present or absent in the gold standard. The numbers of SNVs, MNVs, and InDels were retrieved from the validated and from the initial VCF files (Supplementary Table 3, Supplementary Table 4). Next, true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) were calculated for SNVs, MNVs, and InDels identified by each combination of mapper and variant caller. Finally, performance statistics, such as F1 score, sensitivity, specificity, precision, and accuracy were calculated.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

General stats about mapping of reads

Six Illumina paired-end sequence read datasets[47,48] from *A. thaliana* Nd-1 and one control sample of Col-0 [47] were processed using all combinations of five read mapping and eight different variant calling tools (including three different versions of one tool) to evaluate the mapping percentage as well as precision, sensitivity, and specificity of each variant calling pipeline. Due to these combinations (7×5×10), 350 variant calling sets were generated in this study. Overall, the sequence read quality of the processed data sets was high ranging from an average Phred score of 35 to 38 (Supplementary Table 1). We observed only minor differences between the different sequence read datasets with respect to percentage of properly aligned read pairs (Supplementary Figure 1). In general, a higher proportion of the 2×300 nt paired-end reads was mapped ranging from 94.8% to 99.5%, while the 2×250 nt and the 2×100 nt paired-end reads resulted in mapping proportions ranging from 92.7% to 99.6% and 90.0% to 99.1%, respectively.

The comparison of mapping performance revealed that GEM3 had the highest average percentage of aligned read pairs (99.6%), followed by Novoalign (99.4%), Bowtie2 (99.0%), BWA-MEM (98.7%), and the read mapping function within CLC Genomics Workbench (CLC-mapper) (95.2%) (Figure 2).

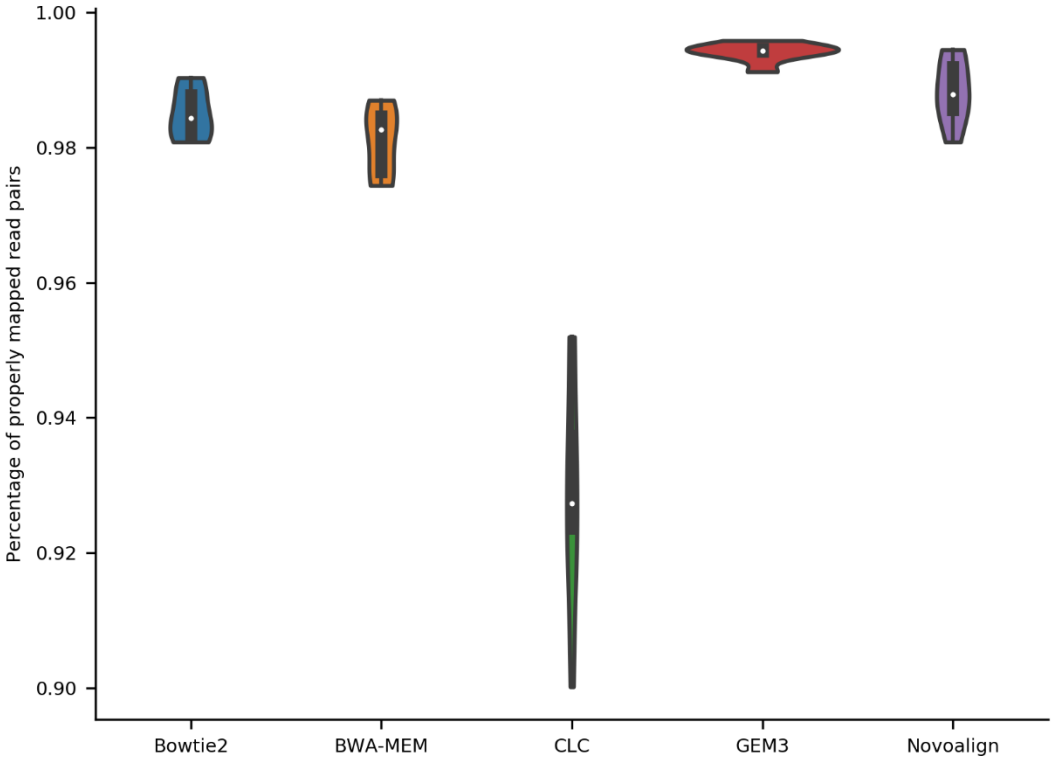

Figure 2: Ratio of mapped sequence read pairs per mapper. Sequence reads of six A. thaliana Nd-1 data sets were mapped to the Col-0 reference genome sequence TAIR10. The average ratio of aligned read pairs was calculated for Bowtie2, BWA-MEM, the mapping function in CLC Genomics Workbench (CLC), GEM3, and Novoalign based on all six data sets through the flagstats function of samtools. The width of the violin plots is proportional to the density of the data points. The boxplots inside the violin plots indicate quantiles and the white dot indicates the median.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Initial variant calling results & validation results

The initial variant calling revealed between 32,939 (Bowtie2/CLC-caller) and 1,009,163 (BWA-MEM/ VarDict) unfiltered SNVs, while the number of unfiltered InDels ranged from 2,559 (BWA-MEM/VarScan) to 240,879 (GEM3/VarDict) (Supplementary Table 3). Based on the three variant callers, which were able to call and classify MNVs (CLC-caller, VarDict, and FreeBayes), MNVs ranged from 1,394 (Bowtie2/CLC-caller) to 168,100 (CLC-mapper/FreeBayes) (Supplementary Table 3).

The quality of a variant call set is determined by the transition/transversion ratio, as a worse variant call set tends to have a lower transition/transversion ratio[49]. While most variant callers showed a similar transition/transversion ratio with a median ranging from 1.256 (LoFreq) to 1.288 (VarDict), SNVer revealed a lower median of 1.2 and especially FreeBayes performed worst, showing a median of 1.15 (Figure 3). In addition, FreeBayes revealed the greatest variation ranging from 0.92 to 1.31.

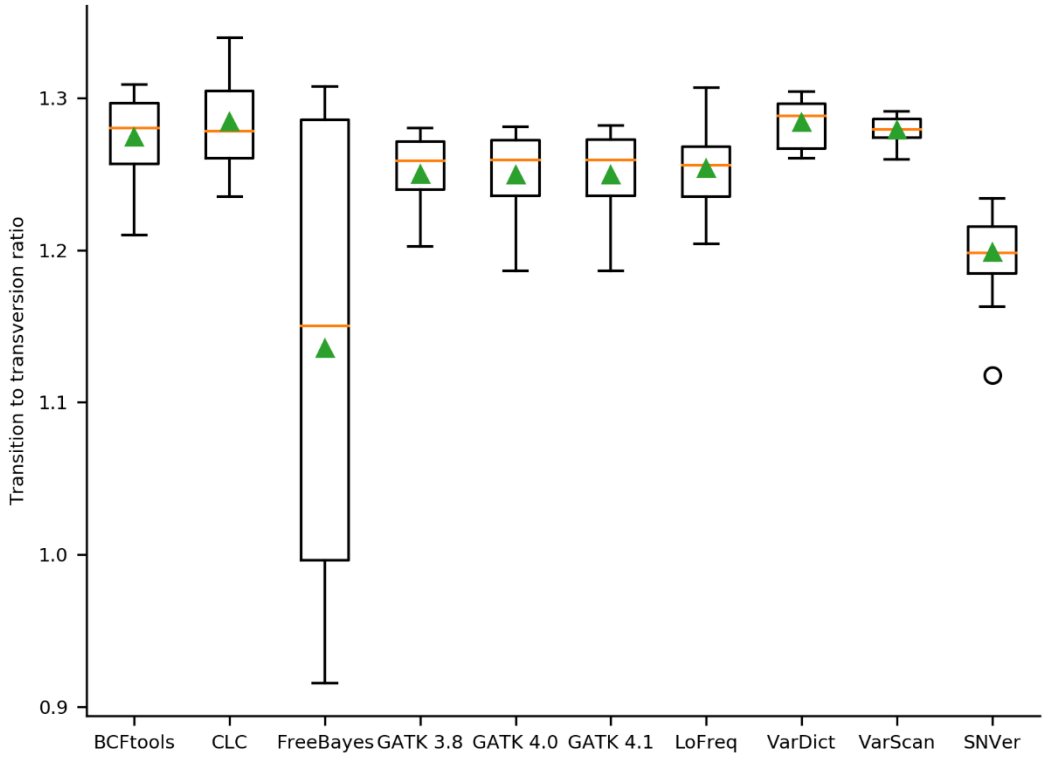

Figure 3: Ratio of transitions/transversions in the variant call sets per variant caller. Evaluation of call set quality was harnessed by analyzing the transition/transversion ratio. The orange line represents the median, the green triangle represents the mean.

In order to analyze whether a variant caller identifies relatively more SNVs than InDels, the ratio of SNVs to SNVs and InDels was calculated per variant caller (Figure 4). BCFtools identified the highest proportion of SNVs (median = 0.90), while VarDict and GATK 4.1 called the lowest proportion of SNVs (median = 0.824). Moreover, all GATK versions performed similar and revealed low variance when

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

compared to the other variant callers. Interestingly, BWA-MEM/VarScan using the SRR3340910 data set yielded the highest SNVs/InDels ratio with almost 1 (0.996).

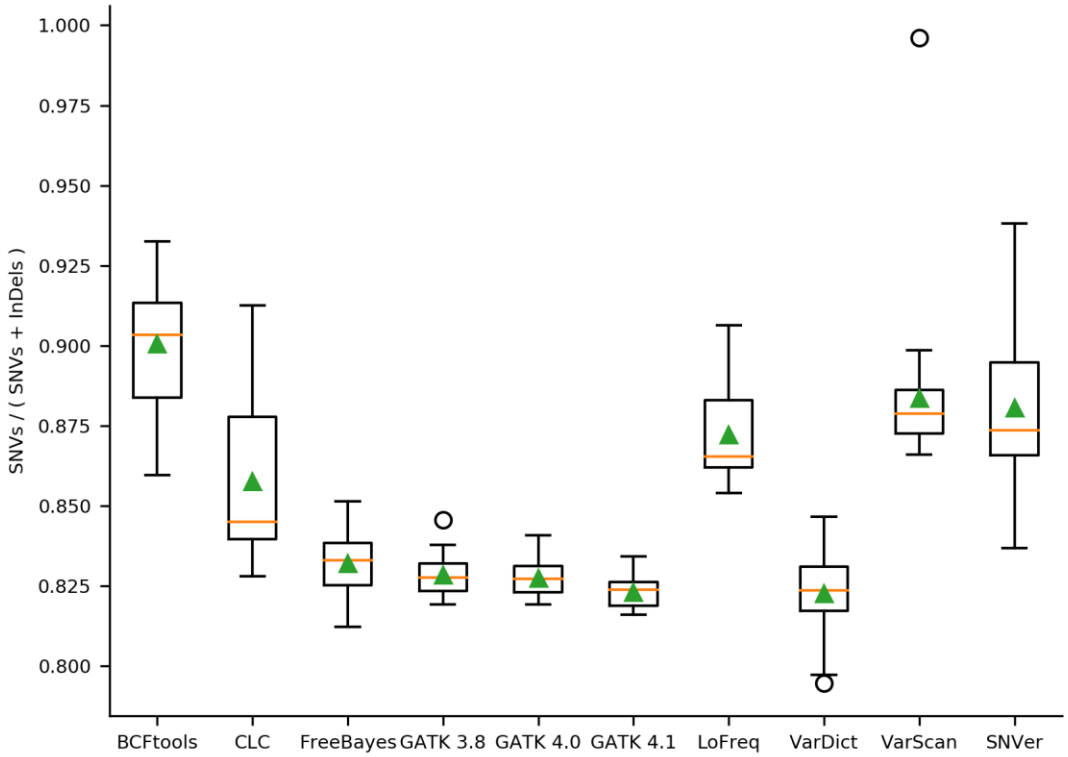

Figure 4: Proportion of SNVs to all variants per variant caller. Performance of each variant caller was assessed based on 30 mappings of A. thaliana Nd-1 reads against the Col-0 reference genome sequence TAIR10. Evaluation of the proportion of SNVs to all variants in the results of each applied variant caller was analyzed. MNPs were excluded because not all variant callers identified MNPs. The orange line represents the median, the green triangle the mean.

To infer whether a variant caller is more prone to call small or large InDels, the distribution of InDel lengths was inspected (Figure 5). Especially VarDict called very large insertions (up to 981 bp) and very large deletions (up to 998 bp), which are likely to be artifacts since they are filtered out in the corresponding validated call set (Supplementary Figure 4). VarScan (134 to -93), SNVer (134 to -95), CLC-caller (156 to -95), LoFreq (168 to -109), and BCFtools (216 to -108) showed a narrower range of InDel lengths.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

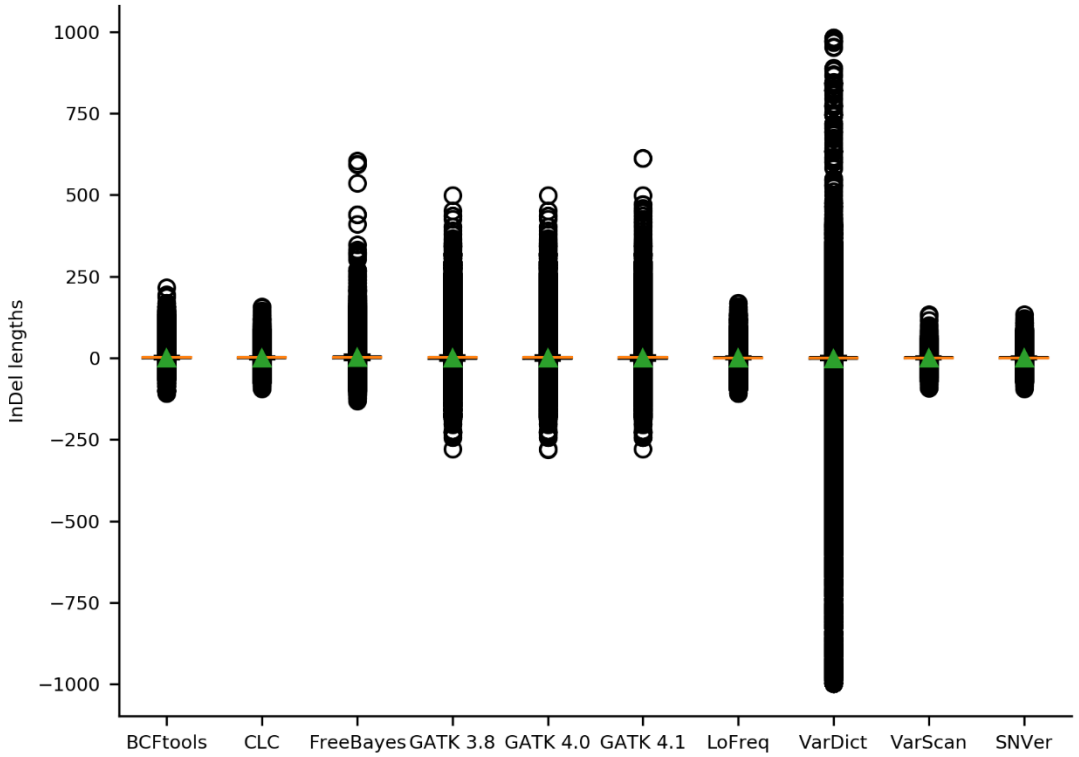

Figure 5: Distribution of InDel lengths per variant caller. Performance of variant callers was assessed based on 30 mappings of A. thaliana Nd-1 reads against the Col-0 reference genome sequence TAIR10. The length distribution of all InDels identified by each applied variant caller was analyzed. The orange line represents the median, the green triangle represents the mean.

In order to compare the performance of different variant calling pipelines, we calculated the sensitivity, specificity, accuracy, precision, and F1-Score (Table 1, Supplementary Table 3). GATK revealed the highest accuracy in combination with most mappers. The only exception is the combination of GEM3 and VarScan, which performed better than any GATK version (Figure 6). GATK worked best on alignments produced by BWA-MEM and Novoalign. All three evaluated GATK versions (v3.8, v4.0, and v4.1) showed an almost identical performance. In general, Novoalign reached the best (median) results with respect to accuracy. The only exceptions are CLC-caller and VarScan based on alignments produced by CLC-mapper and GEM3, respectively. While Bowtie2 was identified to yield high specificity with most variant callers, it showed a low accuracy with most variant callers except for FreeBayes and VarDict.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

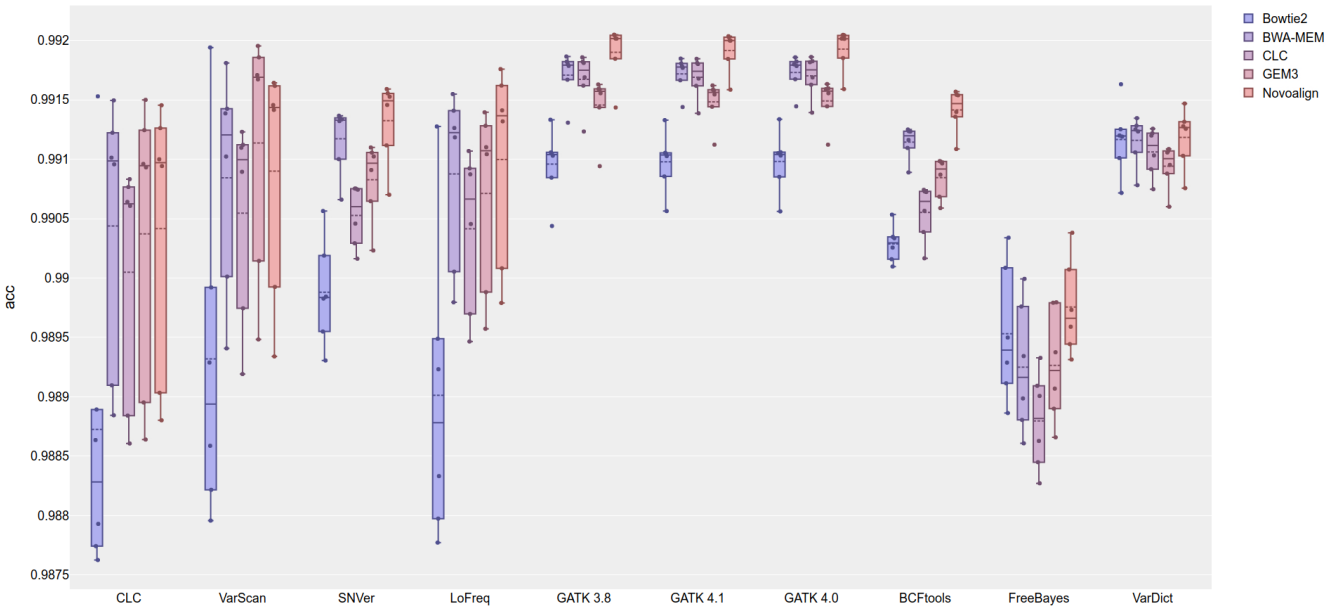

Figure 6: Accuracy of variant calling pipelines. The accuracy for each variant calling pipeline is shown with mean (dashed line) and median (straight line) calculated based on the results of the six analyzed data sets.

In general, the sensitivity of the variant caller pipelines ranged from 0.0219 (Bowtie2/CLC-caller) to 0.6038 (BWA-MEM/VarDict) and the specificity from 0.99450 (CLC-mapper/FreeBayes) to 0.999961 (Bowtie2/CLC-caller) (Supplementary Figure 2, Supplementary Figure 3). Moreover, we observed a negative correlation of -0.8 between specificity and sensitivity, indicating that a pipeline with a high sensitivity showed a low specificity and vice versa. Almost every variant caller, except for VarDict, showed the lowest specificity when used in combination with CLC-mapper, while in parallel these combinations had one of the highest sensitivities. VarDict showed the highest specificity, but lowest sensitivity with Bowtie2 and performed inferior to GEM3 in terms of specificity, while BWA-MEM reached the best results in sensitivity.

All tested GATK versions (v3.8, v4.0, and v4.1) showed a very high sensitivity and were only outperformed by specific VarDict samples, namely the 2×100 nt paired-end data set independent of the choice of the mapper, which reached up to 0.6038 sensitivity (BWA-MEM / VarDict-SRR2919279). However, the specificity of GATK was inferior to some other variant callers. Only minor differences were observed between the three evaluated GATK versions regarding both sensitivity and specificity. The use of different mappers had a substantially higher impact than the applied GATK version.

Followed by GATK, FreeBayes showed a good performance in terms of sensitivity and robust results across all mappers, whereas the other variant callers showed a low performance in combination with Bowtie2. CLC-caller, VarScan, and LoFreq revealed a great variation with respect to sensitivity across all mappers, while GATK and especially VarDict displayed very low variance in their results. When focusing on median sensitivity, the following combinations showed the best results: CLC-mapper/CLC-caller, GEM3/VarScan, CLC-mapper/SNVer, CLC-mapper/LoFreq, CLC-mapper/GATK 3.8, CLC-mapper/GATK 4.0, CLC-mapper/GATK 4.1, CLC-mapper/BCFtools, GEM3/FreeBayes, and BWA-MEM/VarDict. However, in terms of median specificity all variant callers revealed the best results in combination with Bowtie2,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

except for FreeBayes, which worked best with Novoalign. Moreover, FreeBayes showed the lowest performance and largest variation across all mappers (Supplementary Figure 2).

Finally, the harmonic mean of precision and sensitivity, namely the F1 score, was analyzed (Supplementary Figure 6). Novoalign in combination with GATK revealed the best mean performance with respect to the F1 score. Again, different GATK versions showed almost identical performance (Table 1).

Table 1: A Performance statistics of variant calling pipelines. For each variant calling pipeline the statistics to infer performance are listed. sen = sensitivity, spe = specificity, pre = precision, acc = accuracy, F1 = F1 score.

	BCFtools	CLC	FreeBayes	GATK 3.8	GATK 4.0	GATK 4.1	LoFreq	SNVer	VarDict	VarScan
Bowtie2	26.701 sen	7.979 sen	43.729 sen	33.836 sen	33.819 sen	33.869 sen	12.282 sen	21.844 sen	37.343 sen	13.910 sen
	99.928 spe	99.987 spe	99.566 spe	99.937 spe	99.937 spe	99.936 spe	99.985 spe	99.969 spe	99.884 spe	99.979 spe
	81.800 pre	88.244 pre	53.735 pre	87.286 pre	87.342 pre	87.035 pre	91.607 pre	90.107 pre	79.803 pre	89.530 pre
	99.030 acc	98.828 acc	98.939 acc	99.104 acc	99.104 acc	99.104 acc	98.878 acc	98.983 acc	99.120 acc	98.894 acc
	0.403 F1	0.145 F1	0.47 F1	0.488 F1	0.488 F1	0.488 F1	0.214 F1	0.352 F1	0.512 F1	0.239 F1
BWA-MEM	43.186 sen	34.992 sen	43.053 sen	47.130 sen	47.099 sen	47.136 sen	38.874 sen	41.435 sen	40.302 sen	37.506 sen
	99.806 spe	99.908 spe	99.548 spe	99.842 spe	99.842 spe	99.839 spe	99.892 spe	99.868 spe	99.852 spe	99.907 spe
	73.182 pre	81.892 pre	51.289 pre	79.129 pre	79.083 pre	78.786 pre	81.122 pre	79.521 pre	76.686 pre	83.535 pre
	99.120 acc	99.099 acc	98.916 acc	99.179 acc	99.179 acc	99.178 acc	99.122 acc	99.133 acc	99.124 acc	99.120 acc
	0.543 F1	0.492 F1	0.459 F1	0.591 F1	0.59 F1	0.59 F1	0.528 F1	0.546 F1	0.531 F1	0.518 F1
CLC	47.682 sen	36.382 sen	41.564 sen	49.237 sen	49.357 sen	49.413 sen	41.730 sen	45.352 sen	39.508 sen	41.349 sen
	99.695 spe	99.851 spe	99.517 spe	99.817 spe	99.815 spe	99.812 spe	99.821 spe	99.759 spe	99.857 spe	99.847 spe
	65.903 pre	73.878 pre	48.854 pre	77.341 pre	77.183 pre	76.901 pre	73.611 pre	70.493 pre	76.995 pre	76.971 pre
	99.064 acc	99.062 acc	98.882 acc	99.175 acc	99.175 acc	99.174 acc	99.066 acc	99.060 acc	99.112 acc	99.100 acc
	0.549 F1	0.491 F1	0.441 F1	0.601 F1	0.602	0.602 F1	0.536 F1	0.549 F1	0.525 F1	0.539 F1
GEM3	35.488 sen	32.826 sen	44.104 sen	40.895 sen	40.890 sen	40.963 sen	33.867 sen	34.710 sen	39.131 sen	41.502 sen
	99.887 spe	99.932 spe	99.523 spe	99.902 spe	99.902 spe	99.900 spe	99.941 spe	99.925 spe	99.847 spe	99.904 spe
	79.533 pre	85.457 pre	51.481 pre	84.189 pre	84.221 pre	83.863 pre	87.657 pre	85.573 pre	75.648 pre	84.315 pre
	99.092 acc	99.095 acc	98.922 acc	99.157 acc	99.157 acc	99.156 acc	99.107 acc	99.097 acc	99.100 acc	99.169 acc
	0.49 F1	0.475 F1	0.464 F1	0.55 F1	0.551 F1	0.55 F1	0.489 F1	0.494 F1	0.519 F1	0.557 F1
Novoalign	44.160 sen	33.858 sen	42.986 sen	48.599 sen	48.575 sen	48.620 sen	38.892 sen	41.436 sen	39.863 sen	38.170 sen
	99.820 spe	99.922 spe	99.610 spe	99.846 spe	99.845 spe	99.842 spe	99.905 spe	99.885 spe	99.860 spe	99.922 spe
	75.260 pre	84.222 pre	54.799 pre	80.019 pre	79.951 pre	79.643 pre	83.510 pre	81.968 pre	77.296 pre	85.862 pre
	99.147 acc	99.097 acc	98.966 acc	99.202 acc	99.202 acc	99.200 acc	99.137 acc	99.149 acc	99.127 acc	99.144 acc
	0.556 F1	0.483 F1	0.472 F1	0.605 F1	0.604 F1	0.604 F1	0.532 F1	0.551 F1	0.529 F1	0.529 F1

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

The major challenge in many pangenome and re-sequencing projects is the accurate and comprehensive identification of sequence variants. Due to the high diversity and complexity of plant genomes and their differences to animal (e.g. human) genomes, variant callings in plant research differ substantially from those in human and biomedical research. Large amounts and different NGS data types (as reviewed in [11]), the diversity of bioinformatic algorithms, and the quality of the reference genome sequences render the choice of the best approach challenging. Hence, we performed a benchmarking study to provide comparable data showing the performance of combinations of frequently applied mappers and variant callers (variant calling pipelines) on plant data sets. A previous report[44] is extended by providing data about the performance of additional tools both for the mapping and variant calling step.

To allow for a consistent comparison of baseline performance, we used default parameters for all tools as these parameters are frequently chosen in plant science applications[4,5,9]. Sequence read data sets from different sequencing platforms, with different read lengths, and different sizes were processed to ensure a realistic benchmarking of tools. Since all evaluated tools can process a real plant data set within a day, we refrained from assessing the computational costs of the analysis. There is usually a trade-off between quality of the results and computational costs. In our experience, many plant scientists select the workflow leading to the best results independent of computational costs[50].

The first step in a variant calling pipeline is the alignment (mapping) of reads to a reference sequence. While the mapping of 2×250 nt paired-reads resulted in a higher mapping percentage, the performance difference to 2×100 nt reads is only about 10%. As different sequencing platforms were used for the data generation, per base quality might contribute to this difference. It is expected that longer reads are aligned with higher specificity and hence improve the following variant calling.
The quality of the variant calling sets was assessed by the transition/transversion (ti/tv) ratio which was previously described as a quality indicator[49]. Overall the quality of almost all analyzed call sets was similar. A previous benchmarking study with SAMtools and GATK reported similar ti/tv ratios for all pipelines[51]. A filtering step increased the ti/tv ratios, indicating that the filtering increased the quality of the call sets[51]. This observation is in line with our findings, which revealed an increased ti/tv for the filtered call sets reduced to variants present in the gold standard (Supplementary Figure 5, Supplementary Table 4). As FreeBayes showed a substantial increase in the quality through filtering, we recommend checking the transition/transversion ratio when applying FreeBayes. This effect might be dataset specific.

The choice of the variant caller is crucial for the number of called SNVs, MNVs, and InDels. For example, only CLC-caller, VarDict and FreeBayes were able to call MNVs, thus being more suitable for the identification of structural differences. Furthermore, variant caller results differ with respect to the ratio of SNVs to InDels, which should be considered depending on the specific requirements of the respective sequencing project. BCFtools called relatively more SNVs than InDels, while GATK revealed relatively

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

more InDels. A unique property of VarDict was the detection of InDels up to almost 1 kb which exceeds the read length. Since the accurate identification of such large variants, which are longer than the average read length, is still a challenging task[52], many of these variants are likely false positives. Moreover, the reduced amount of large insertions in the validated call sets of VarDict supports this assumption.

Depending on the application, a pipeline with a high sensitivity or high specificity is desired. In terms of sensitivity, GATK in combination with CLC-mapper, Novoalign and BWA-MEM yielded the best and most consistent results across all evaluated data sets. These results are in line with a recent study showing that GATK often outperformed SAMtools in terms of sensitivity, precision, and called raw InDels[44]. A high sensitivity is essential when a high number of true positives variations accelerates the power of the analyses, e.g. when looking for a detrimental variation between two samples. In this study, a pipeline comprising Bowtie2 and LoFreq resulted in the highest specificity and can thus be recommended. In contrast, a high specificity is desired in mapping-by-sequencing (MBS) projects, as a high proportion of true positives can keep the signal to noise ratio high. Combining both performance metrics by analyzing the accuracy, best results were obtained with Novoalign and GATK. The same pipeline yielded the best results regarding the harmonic mean of precision and sensitivity (F1 score). Differences observed between the three evaluated GATK versions (v3.8, v4.0, and v4.1) were negligible. However, functionalities and computational performance might differ between these versions.

In summary, this benchmarking study provides insights into the strengths and weaknesses of different variant calling pipelines when applied on plant NGS data sets. Although the performance of all evaluated tools will differ between samples depending on properties of the read datasets and the genome sequence, we hope that our findings serve as a helpful guide.

Acknowledgments

We thank the Bioinformatics Resource Facility support team of the CeBiTec for great technical support. We acknowledge support for the Article Processing Charge by the Open Access Publication Fund of Bielefeld University.

Author Contributions

H.M.S., A.R., and B.P. conceived the project. H.M.S., A.R., and B.P. conducted data analysis. H.M.S., B.P. wrote the manuscript. B.P. supervised the project. All authors read and approved the final version of the manuscript.

References

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. Weigel, D. & Mott, R. The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biology* **10**, 107 (2009).

2. Varshney, R. K., Pandey, M. K. & Chitikineni, A. *Plant Genetics and Molecular Biology*. (Springer, 2018).

3. Li, Y. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**, 1045–1052 (2014).

4. Zhao, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**, 278–284 (2018).

5. Song, J.-M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* **6**, 34–45 (2020).

6. Abe, A. *et al.* Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol* **30**, 174–178 (2012).

7. Liu, S., Yeh, C.-T., Tang, H. M., Nettleton, D. & Schnable, P. S. Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq). *PLOS ONE* **7**, e36406 (2012).

8. Mascher, M. *et al.* Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biology* **15**, R78 (2014).

9. Ries, D., Holtgräwe, D., Viehöver, P. & Weisshaar, B. Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* **17**, 236 (2016).

10. Pfeifer, S. P. From next-generation resequencing reads to a high-quality variant data set. *Heredity* **118**, 111–124 (2017).

11. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32 (2009).

12. Andrews, S. FastQC: A quality control tool for high throughput sequence data. [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (2010).

13. Planet, E., Attolini, C. S.-O., Reina, O., Flores, O. & Rossell, D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**, 589–590 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Dai, M. *et al.* NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* **11 Suppl 4**, S7 (2010).

15. Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* **27**, 130–131 (2011).

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

17. Rodríguez-Ezpeleta, N., Hackenberg, M. & Aransay, A. M. *Bioinformatics for High Throughput Sequencing*. (Springer Science & Business Media, 2011).

18. Reinert, K., Langmead, B., Weese, D. & Evers, D. J. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics* **16**, 133–151 (2015).

19. Shang, J. *et al.* Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. *BioMed Research International* https://www.hindawi.com/journals/bmri/2014/309650/ (2014) doi:https://doi.org/10.1155/2014/309650.

20. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* **15**, 256–278 (2014).

21. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

22. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).

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

24. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).

25. Ruffalo, M., LaFramboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790–2796 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Yu, X. *et al.* How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Min* **5**, 6 (2012).

27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

28. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]* (2012).

29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

30. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

31. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-11.10.33 (2013).

32. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018) doi:10.1101/201178.

33. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40**, 11189–11201 (2012).

34. Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* **39**, e132 (2011).

35. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).

36. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

37. Pucker, B. & Schilbert, H. Genomics and Transcriptomics Advances in Plant Sciences. in *Molecular Approaches in Plant Biology and Environmental Challenges* (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

38. Fumagalli, M. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE* **8**, e79667 (2013).

39. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).

40. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep* **5**, (2015).

41. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE* **11**, e0151664 (2016).

42. Bian, X. *et al.* Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19**, (2018).

43. Hwang, K.-B. *et al.* Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Sci Rep* **9**, 1–10 (2019).

44. Wu, X., Heffelfinger, C., Zhao, H. & Dellaporta, S. L. Benchmarking variant identification tools for plant diversity discovery. *BMC Genomics* **20**, 701 (2019).

45. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202-1210 (2012).

46. Baasner, J.-S., Howard, D. & Pucker, B. Influence of neighboring small sequence variants on functional impact prediction. *bioRxiv* 596718 (2019) doi:10.1101/596718.

47. Kleinboelting, N. *et al.* The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. *Mol Plant* **8**, 1651–1664 (2015).

48. Pucker, B. *et al.* A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. *PLoS ONE* **11**, e0164321 (2016).

49. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.986059; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

50. Pucker, B. *et al.* A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLOS ONE* **14**, e0216233 (2019).

51. Liu, X., Han, S., Wang, Z., Gelernter, J. & Yang, B.-Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLOS ONE* **8**, e75619 (2013).

52. Kavak, P. *et al.* Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics* **33**, i161–i169 (2017).