Genotyping of evolving prokaryotic populations

Markus Zojer [1] , **Lisa N Schuster** [2] , **Frederik Schulz** [2] , **Alexander Pfundner** [1] , **Matthias Horn** [2] , **Thomas Rattei** Corresp. [1]

[1] Department of Microbiology and Ecosystems Science, Division of Computational Systems Biology, University of Vienna, Vienna, Austria
[2] Department of Microbiology and Ecosystems Science, Division of Microbial Ecology, University of Vienna, Vienna, Austria

Corresponding Author: Thomas Rattei
Email address: thomas.rattei@univie.ac.at

Genomic heterogeneity of bacterial species is observed and studied in experimental evolution experiments, clinical diagnostics and occurs as micro-diversity of natural habitats. The challenge for genome research is to accurately capture this heterogeneity with the currently used short sequencing reads. Recent advances in NGS technologies improved the speed and coverage and thus allowed for deep sequencing of bacterial populations. This facilitates the quantitative assessment of genomic heterogeneity, including low frequent alleles or haplotypes. However, false positive variant predictions due to sequencing errors and mapping artifacts of short reads need to be prevented. We therefore created VarCap, a workflow for the reliable prediction of different types of variants even at low frequencies. In order to predict SNPs, indels and structural variations, we evaluated the sensitivity and accuracy of different software tools using synthetic read data. The results suggested that the best sensitivity could be reached by a combination of different tools. We identified possible reasons for false predictions and used this knowledge to improve the accuracy by post-filtering the predicted variants according to properties such as frequency, coverage, genomic environment/localization and co-localization with other variants. This resulted in the reliable prediction of variants above a minimum relative abundance of 2%. VarCap is designed for being routinely used within experimental evolution experiments or for clinical diagnostics. The detected variants are reported as frequencies within a vcf file and as a graphical overview of the distribution of the different variant/allele/haplotype frequencies. The source code of VarCap is available at https://github.com/ma2o/VarCap. In order to provide this workflow to a broad community, we implemeted VarCap on a Galaxy webserver (Afgan et al. 2016) , which is accessible at http://galaxy.csb.univie.ac.at.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Genotyping of evolving prokaryotic populations

Markus Zojer[1,2,*], Lisa N. Schuster[2], Frederik Schulz[2], Alexander Pfundner[1], Matthias Horn[2] and Thomas Rattei[1]

[1]Department of Microbiology and Ecosystems Science, Division of Computational Systems Biology, Althanstrasse 14, 1090 Wien

[2]Department of Microbiology and Ecosystems Science, Division of Microbial Ecology, Althanstrasse 14, 1090 Wien

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

ABSTRACT

Genomic heterogeneity of bacterial species is observed and studied in experimental evolution experiments, clinical diagnostics and occurs as micro-diversity of natural habitats. The challenge for genome research is to accurately capture this heterogeneity with the currently used short sequencing reads. Recent advances in NGS technologies improved the speed and coverage and thus allowed for deep sequencing of bacterial populations. This facilitates the quantitative assessment of genomic heterogeneity, including low frequent alleles or haplotypes. However, false positive variant predictions due to sequencing errors and mapping artifacts of short reads need to be prevented.

We therefore created VarCap, a workflow for the reliable prediction of different types of variants even at low frequencies. In order to predict SNPs, indels and structural variations, we evaluated the sensitivity and accuracy of different software tools using synthetic read data. The results suggested that the best sensitivity could be reached by a combination of different tools. We identified possible reasons for false predictions and used this knowledge to improve the accuracy by post-filtering the predicted variants according to properties such as frequency, coverage, genomic environment/localization and co-localization with other variants. This resulted in the reliable prediction of variants above a minimum relative abundance of 2%.

VarCap is designed for being routinely used within experimental evolution experiments or for clinical diagnostics. The detected variants are reported as frequencies within a vcf file and as a graphical overview of the distribution of the different variant/allele/haplotype frequencies. The source code of VarCap is available at https://github.com/ma2o/VarCap. In order to provide this workflow to a broad community, we implemeted VarCap on a Galaxy webserver (Afgan et al. 2016), which is accessible at http://galaxy.csb.univie.ac.at.

1. INTRODUCTION

The genotyping of heterogeneous populations of one prokaryotic species is an increasingly important method to address microbiological questions regarding population composition and dynamics under prevalent selective pressures. This approach is e.g. used in experimental evolution experiments (Barrick and Lenski 2013) and studies of host - pathogen systems (Gardy et al. 2011, Bos et al. 2011, McElroy, Thomas, and Luciani 2014). Recent developments in Next-Generation-Sequencing (NGS) technologies allow for sequencing at high coverage within a short timeframe, however limited to short read length.

The classical approach of assembling genomes out of short DNA reads preferably reconstructs the most abundant genotype into genome contigs and scaffolds. In order to retrieve haplotype frequency information, reads need to be mapped onto the assembly or a reference genome. Variant calling is then performed on the alignment of the reads. The predicted variants can be phased into haplotypes or alleles if a whole haplotype reconstruction is not possible due to insufficient linkage of the variant sites. The variant prediction, however, can lead to false positives due to sequencing errors, such as indels and substitutions. The reads may be misplaced during mapping due to their short length and thus can lead to false positive variant calls (Li 2014). Sequencing errors can be partially reduced by quality filtering and error correction (Yang, Chockalingam, and Aluru 2013). As a consequence, the substitution error rate for Illumina could be decreased below one percent while indel homopolymer errors showed to accumulate logarithmically with the length of the stretches (Minoche, Dohm, and Himmelbauer 2011) and can thereby reliably identified.

Most of the genotyping studies of prokaryotes so far have been done by resequencing of clonal bacterial cultures (Maharjan et al. 2013, Blount et al. 2012). Deep sequencing of non-clonal cultures was mainly done for metagenomic profiling of communities (Qin et al. 2010) and only to a minor extend for the characterization of allele frequencies (Eyre et al. 2013). The genotyping of non-clonal variants in heterogeneous populations, however, remains challenging (DePristo et al. 2011, Nielsen et al. 2011, Kofler and Schlötterer 2014, Pulido-Tamayo et al. 2015).

In order to get a most complete picture of the different haplotype or allele frequencies it is fundamental to exploit high coverage sequencing data and to detect all types of variants, which are SNPs, Indels and structural variations (SV). Therefore it is necessary to integrate several variant calling software tools, which utilize different approaches for the detection of the different kinds of variants.

Commonly used tools to identify SNPs are Samtools/bcftools and GATK (Li et al. 2009, McKenna et al. 2010). These tools were developed with the assumption to detect variants within diploid organisms, which limits their detection power for haploid prokaryotes. Therefore we also considered the more generic tool VarScan2 (Koboldt et al. 2012), which can predict SNP frequencies in low and high coverage data, and LoFreq-Star (Wilm et al. 2012), which was specifically developed to predict SNPs at low frequencies in high coverage datasets. They all work on read alignments or mpileup files and use read and mapping quality scores as well as strand bias filters to reliably detect SNPs. In addition Samtools/bcftools and VarScan2 can also be used to identify small indels. Pindel (Ye et al. 2009) uses a pattern growth algorithm to detect small and large indels from 1bp up to 10kb. Large indels and structural variations (SV), such as translocations, duplications and inversions, are detected by Breakdancer and Delly (K. Chen et al. 2009, Rausch et al. 2012), as they make use of insert size deviations, paired end information and split read information to find variations larger than 300 bp. As an alternative, Cortex_var (Iqbal et al. 2012) does not rely on mapped reads but uses *de novo* assembled contigs, which are compared to each other or to a reference in order to identify most kinds of variants. All those approaches have been designed for different degrees of heterozygosity, ranging from diploid genomes to multiploid populations with low abundant genotypes.

The genotyping of prokaryote populations in experimental evolution experiments is typically based on many NGS datasets with high coverage. There is therefore a demand for fully automated software for read mapping and variant calling, which is both sensitive and accurate, aware of low abundant subpopulations, and which considers all possible types of variants. To the best of our knowledge no such software workflow has been published so far. In this study we have evaluated variant callers on synthetic data in order to determine and compare their sensitivity and accuracy. This allowed us to develop and validate VarCap, a workflow for accurate and sensitive genotyping of prokaryotic populations. We finally applied VarCap to a long-term experimental evolution experiment of a bacterial symbiont of amoebae.

2. METHODS

2.1 Creating synthetic variant genomes: Ideally, the organism selected for simulation should exhibit generic properties that make the results applicable for most prokaryotes. In our simulation and evaluation of the variant detection prototype, however, we decided to pick the non-model organism *Protochlamydia amoebophila.* It offered the unique opportunity to experimentally validate variant predictions immediately during the software development. In addition, *P. amoebophila* exhibits typical properties as its genome size of 2.4 Mb is close to one of the main peaks in the bacterial and archaeal genome size distribution (Koonin and Wolf 2008). For validation purposes we additionally used 6 different organisms that we selected to represent the diversity of prokaryotic genomes regarding G+C contents and genome size.

Variant datasets were created by randomly inserting different types of variants into reference genomes downloaded from the NCBI Refseq database (Pruitt et al. 2012)(**Supp. Table 1**). We created mixed types of datasets containing 135 variations, as well as datasets containing one specific type of variant. The 135 variants of the mixed type dataset consisted of 100 SNPs, 10 small Indels, 10 large Indels and 5 translocations, 5 duplications (including one double duplication) and 5 inversions (Set: mixvar_1-7). We used a SNP/Indel ratio of 10 for small Indels and 20 for all Indels as SNP/Indel ratios for bacterial genomes was reported between 15 and 20 (Moran, McLaughlin, and Sorek 2009, J.-Q. Chen et al. 2009). Large insertions hereby also mimic the process of horizontal gene transfer. As structural variations are reported to be crucial for bacterial genome evolution, we also added few translocation, duplication and inversion sites to challenge the detection software.

The SNPs were positioned as single seeds, to which the remaining SNPs were randomly assigned within a distance of 1 to 15 bases (max 4 allowed) with decreasing probability in order to create mutation hotspots. The size of the large Indels was randomly chosen between 5 and 2000 nucleotides, while the size of translocations, duplications and inversions varied from 300 to 2000 nucleotides. The datasets harboring only one type of variant contained either 100 SNPs, 100 small Indels, 100 large Indels, 50 translocations, 50 duplications or 50 inversions (Sets mono_02-07).

ALFSim was used (Dalquen et al. 2012) to simulate genome evolution and to generate more distantly evolved subpopulations. Therefore coding and intergenic nucleotide sequences according to the genome annotation were extracted, which served as input for ALFSim. We selected a simulated subspecies having a nucleotide dissimilarity of 0.8% resulting in 21000 SNPs, 100 Indels and 3 duplications.

2.2 Sequencing read simulation: We used SimSeq (https://github.com/jstjohn/SimSeq, version from 4.12.2011, Earl et al. 2011) and pIRS (pIRS_1.10, Hu et al. 2012) for the simulation of 100 nucleotides (nt) paired end Illumina reads with an insert size of 250 nt and an insert size standard deviation of 10, 20 and 30%. For pIRS we used the supplied error model, while for SimSeq the updated empirical error models for forward and reverse strand were used (hiseq_mito_default _bwa_mapping_mq10_1_Corrected.txt, hiseq_mito_default _bwa_mapping_mq10_2_Corrected.txt).

2.3 Sequence read processing and mapping: Read quality was determined using FastQC (v0.10.0, Patel and Jain 2012). The quality filtering and trimming of sequencing reads was done by Prinseq-lite (0.19.5, Schmieder and Edwards 2011) and Trimmomatic (0.32, Bolger, Lohse, and Usadel 2014) and applied with the following settings: First a sliding window with size 10 removed any bases with lower quality than 20 from the 3' side. Then we filtered out all reads shorter than 40 nt and finally we discarded any read with an average Phred score below 30. Only read pairs were kept. These reads were mapped against the reference genome using bwa-mem (bwa-0.7.5a, unpublished, Li and Durbin 2009) with standard settings and stored as bam files. For conversions from sam to bam files and from bam to fastq files (as Cortex_var input) we used samtools (0.1.18, Li et al. 2009) and picard-tools (v1.92, http://picard.sourceforge.net/).

Mapping artifacts: In order to emulate mis-mapped reads due to an incomplete reference genome, we mapped reads that were generated from an updated (newly assembled) reference genome back to the older and about 20kB shorter version and to the current version. This dataset did not contain any simulated variants.

2.4 Variant calling: In order to assess true and false positive variant detection rates, artificial non-clonal populations containing SNPs, Indels and SV at abundances of 40%, 20%, 10% 5% and 2% were simulated. We used samtools/bcftools (0.1.18, Li et al. 2009), GATK-lite (GenomeAnalysisTKLite-2.2-8, McKenna et al. 2010), VarScan2 (2.3.6, Koboldt et al. 2012), LoFreq (0.6.1, Wilm et al. 2012) and LoFreq2 (lofreq-star 2.0.0 beta 1, unpublished). For the detection of small Indels we used VarScan2 and pindel (024t, Ye et al. 2009). For large Indels and structural variations (SV) we used pindel, which is described to work well between on variations between 1 and 1000 nt, breakdancer (breakdancer-1.1_2011_02_21, K. Chen et al. 2009) and delly (0.0.11, Rausch et al. 2012) (both start calling SV at 300 nt). Additionally we used the assembler cortex_var (CORTEX_release_v1.0.5.14, Iqbal et al. 2012), which can detect variations by comparing assembled contigs to a reference genome sequence. The sensitivity and precision of the combined workflow were calculated as: sensitivity = TP / (TP+FN), and precision = TP / (TP + FP).

2.5 Setting the minimum abundance for a variant: In order to call a variant it has to be present within a minimum count of sequencing reads. Some variant callers need a variant to be present on 4-8 reads, so we set 8 reads as the minimum absolute abundance (MAA). However, as read coverage slightly varies along the genome, we also used minimum relative abundance (MRA), which is the percentage of variant reads compared to the total coverage. So a MAA of 8 reads corresponds to a MRA of 2% at 400x total coverage.

2.6 Examining the similarity of repetitive regions: We used the edit distance in order to measure the similarity of repetitive regions. The edit distance measures the similarity of 2 sequences by counting the differences between them. This difference can be a substitution, insertion or deletion of a nucleotide. Therefore a edit distance of one means, that two sequences differ in either a substitution, insertion or deletion of a nucleotide.

2.7 Analysis of a long-term experimental evolution experiment: We applied the VarCap workflow to a long-term experimental evolution experiment in order to evaluate its performance on Illumina PE data. Two independent laboratory cultures of the amoeba symbiont *Protochlamydia amoebophila* were subjected to NGS sequencing using the Illumina Genome Analyzer II platform (100bp PE reads, 250bp insert size, 3000x coverage, 250bp insert size) about nine years after its genome was initially sequenced by Sanger sequencing (Horn et al. 2004). For analysis, the obtained Illumina reads were randomly split into replicate read packages with 250 fold coverage each and utilized to detect variant sub-populations at different abundances.

2.8 PCR verification of variations: To verify the variations at positions 1339224, 1339720, and 1338568 in the genome of *P. amoebophila* we amplified the region 1338371-1339843 by PCR using the primers LS0003 5'-AGCTGCATCATTTATCTTCTAG-3' and LS0004 5'-ATCAGTCCACCTACTATCATG-3'. The obtained 1472 bp fragment was cloned into the pCR4-TOPO vector (Invitrogen). Subsequently, 16 of the obtained colonies were picked, and the presence of variations in the cloned amplicons was checked. Clones were sequenced by Sanger sequencing with the primers T3 and T7. Similarly, 14 putative variations in a repetitive region between positions 1533689 and 1534636 were assessed using the primer pair LS0005 5'-TCTCTAGCTCTTTCGCAAATTG-3' and LS0006 5'-CAGTGTTTAACTGGCTGAAAC-3'.

2.9 A Galaxy instance of VarCap

We simplified the use of VarCap for non-experts to a 3 step process facilitated by our Galaxy server (Afgan et al. 2016): I) Create account and login, II) Upload your data to Galaxy and III) Run the VarCap workflow. After the workflow has finished, the user is informed via Email notification. The results are viewable at and downloadable from the website. The output files consist of a vcf file with a detailed description of the variants as wells as two pdf files, which contain overview information about variant and total coverage and frequency information.

3. RESULTS

3.1 Determination of methods capable of sensitive detection of low abundant variations

3.1.0 Evaluation strategy

At the moment there is no software tool or method, that could detect all different types of variants simultaneously, which are relevant for prokaryotic genomes. Therefore we separately evaluated variant detection tools for SNPs, InDels and structural variants (SV). Representative methods for these three targets were selected according to their underlying methodologies. In order to identify the variant calling tools that most sensitively and reliably detect low abundant variant, we initially utilized our most basic variation model (mixvar_1). It incorporates examples of the typical and expected types of variations in microbial genomes, located in typical distances to each other. From these results we constructed a preliminary software framework, which was used as basis for the further evaluations and improvements.

3.1.1 SNPs

Among the many available SNP calling software tools we have compared LoFreq-Star, Varscan2, GATK, Samtools/bcftools and Cortex_var. All of these tools, except Cortex_var, rely on the mapping of reads to a known reference. Cortex_var, instead, *de novo* assembles variant reads into contigs and thereby detects SNPs. Samtools/bcftools and GATK were only designed for homozygous and heterozygous genomes (Yost et al. 2013), whereas LoFreq-Star, Varscan2 and Cortex_var should be able to detect low frequency variants from high coverage sequencing data. Variants were simulated and evaluated at minimum relative abundance (MRA) cutoffs of 10%, 5% and 2%. This means that ideally all variants present at and above those frequencies should be detected. At a MRA of 10%, variants were detected by all SNP calling software tools at a similar sensitivity (**Fig. 1 A**). According to the expectations, the detection rate of GATK and Samtools/bcftools was worse compared to the other programs when the MRA was reduced to 5%, 2% and 1% (**Fig. 1 A**). At a low MRA of 1% LoFreq-Star shows less sensitivity than Varscan2. This is to be expected, as LoFreq-Star builds its own error model and detection threshold to avoid FP and therefore detects no variants below that threshold **Fig. 1 A**).

3.1.2 Indels

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Varscan2 and Pindel were used for the detection of small Indels, and Pindel, Breakdancer, Delly and Cortex_var for the detection of larger Indels. For small Indels, the MSA approach used by Varscan2 should perform at a similar rate as the pattern growth algorithm used by Pindel. Pindel, however, is designed to detect indels from 1-10000bp as it uses a mapping/pattern growth/split read approach. Therefore it should be able to detect the positions of small and large indels with base pair precision. Breakdancer and Delly are designed for the detection of Indels larger than 300bp. They use paired end read information for Indel detection, therefore the position of the large Indels may not be reported at bp resolution. Cortex_var is expected to be less sensitive because of the *de-novo* assembly approach, however it can supply more information than the mapping approaches, including e.g. position, length and sequence of an insertion.

The detection rate of Indels showed little effect to different MRA values (**Fig. 1 B**) (except Samtools/bcftools, see discussion above). Instead, the sensitivity is related to the methodology underlying the software. We observed that Varscan2 can only detect very short Indels (1bp) with the same sensitivity as Pindel, which detected all sizes of Indels with high precission. According to our expectations Breakdancer should have a diminished detection rate for large insertions, as it only considers information about insert size deviation of paired reads and regions with an increased number of anomalous read pairs. We found, that it detects 100% of all large deletions but misses all insertions. As expected, the assembly method used by Cortex_var performs inferior compared to the others. However, it was one of the only two tools that were able to detect large insertions. It detected one third of the large insertions and reported the inserted sequence, while pindel detected the position of large insertions at a higher rate, but without revealing any details.

3.1.3 Structural variations (SV)

For the detection of SV we used Pindel, Breakdancer and Delly, and we added Cortex_var specifically for inversions. They differ slightly in their methodological approaches, therefore we expected Delly to be superior to Breakdancer because of the additional split read alignment. Moreover we expect a limitation of Pindel at larger rearrangements, because the pattern growth algorithm is used within defined limits (up to 10kb). All tools should be able to detect inversions, however they are reported as being harder to detect than other SVs. Breakdancer and Delly detected SV, like duplications and transpositions, regardless of the MRA with high sensitivity (>90%). As expected, the detection rate of Pindel is lagging behind (80%) according to of the suggested internal limits of 10kb. However, the pattern growth method of Pindel was more precise in terms of position and length of the SV as it always hit the exact starting position while Breakdancer and Delly can be off up to 70 bases (**Fig. 1 C**). We additionally found, that large Indels were called at the sites of translocations events (**Fig. 1 C**). This is not entirely unexpected, as a translocation consists of an excision and the consecutive insertion of the excised genomic fragment. The excision can also be seen as a deletion of a fragment and is therefore a partial detection of a more complex type of variant.

Inversions, however, could only be detected at a minor fraction as break positions by Pindel (70% as break positions) and as inversion by Cortex_var (10%) (**Fig. 1 C inv**).

3.1.4 Selected software tools for VarCap

We use LoFreq-Star and Varscan2 for SNPs and Varscan2 and Pindel for small Indels for composing VarCap because they showed similar sensitivity although using different methodological approaches. For larger variants or SV we observed that a combination of pattern growth, split read and paired end read information approaches, which are used by Pindel, results in high sensitivity. This method works well within defined limits (1-10kb). By using only paired end information (Breakdancer), it is possible to detect larger variants at the cost of a lower length limit (300bp) and a coarser resolution of the variant position. Cortex_var, however, was inferior in sensitivity but revealed more information about the detected variants by using a *de-novo* approach. This information can be used to correctly identify the type, position, length or sequence of the variant. Therefore we use Pindel, Breakdancer and Cortex_var for large Indels and Breakdancer, Delly, Pindel and Cortex_var for SV.

Due to the different variant calling abilities of the different tools at low frequencies, we combined different tools to increase the sensitivity (**Fig. 2 A**). Beyond sensitivity we also monitored the precision of the different tools for each type of variant in order to avoid methods that have excessive numbers of FP (**Fig. 3**). As a consequence, Cortex_var was used to predict InDels and inversions but not for SNPs as it accumulated many false positive SNPs in certain areas at low frequencies. Taking together all selected software tools we were able to detect all variants, except inversions, at a MRA of down to 2% with high sensitivity (**Fig. 4**).

3.2 VarCap – a variant calling workflow with high sensitivity and specificity

3.2.1 False Positives due to sequencing errors

False positives occur due to sequencing errors, which are typically present at and below a rate of 1%, therefore we expect them to cause FP calls at and below this relative abundance. In order to study the influence of sequencing errors on different software detection tools, we analyzed 7 differentially composed samples at MRAs of 2% and 1% (mono_02-07). At a MRA of 2% we observed a false positive rate for SNPs, small Indels and Duplications of 0.5 to 1 FP per Megabase (Mb) (**Fig. 5 B MRA 2**). At a lower MRA of 1%, we observed an increase in FP (**Table 1**). At a MRA of 1%, we could nearly completely find all types of variants, except inversions, which we could identify at a rate of 95%. However, the false positive rate for SNPs increased to 80 FP per Mb, while the FP rate for other variants stayed below one FP per Mb (**Fig. 5 A,B MRA 1**). This clearly demonstrates that false positive SNPs are caused by sequencing errors, while the other types of variants stayed at the low rate (~1FP/Mb).
In order to get more insights about the other FP, we examined them in detail at both MRAs. We found that FP of small Indels locate within repetitive regions of the genome. These regions are almost identical areas of the genome at a size that is longer than the insert size of the reads and have an edit distance of 3 or less. Due to their similarity, variant reads can be mapped to similar regions and cause FP calls there.
In order to evaluate how MAA and coverage influence the FP rate, we simulated sequencing coverage from 80 to 1600x (using mixvar_1) and used MAAs from 4 to 20 to filter FP (**Fig. 6**). We detected, that it is necessary to use an MAA cutoff in addition to an MRA cutoff to avoid FP calls at lower coverages (**Fig. 6,** see MRA2 at coverage 160x).

In order to identify and exclude false positives we apply the following filters: To avoid FP SNP calls caused by sequencing errors we apply a MRA of 2%. To avoid FP due to reads mapped to repetitive regions, we mask nearly identical regions according to the properties described above within the reference genome and tag variants that are found within these regions. In order to resolve FP that are caused by incomplete detection of the true variant type, we prioritize larger over smaller variants. Thereby we assign smaller variants to larger ones, if they describe a component of the whole variation e.g. large Indel at excision site of translocation.

3.2.2 FP due to mismapped reads

Mismapped reads have been reported as the cause of FP (Li 2014). Thereby incomplete reference genomes lead to reads getting mapped to similar regions and cause FP calls there. To review this finding at a MRA of 2%, we mapped reads without variants back onto an artificially shortened reference genome. We observed ~180 FP SNPs/75 FP per Mb which were present at different abundances (20%, 8%, 3%) and grouped into hotspots (**Fig. 7 A**). False positive variants were not observed when mapping the reads to the correct reference (**Fig. 7 B**). This finding strongly supports our assumption that wrongly mapped reads cause FP variant calls. A closer investigation of the relevant regions revealed the presence of neighboring break positions, which may indicate both: either a larger structural variation or mismapped reads due to an incomplete reference genome.

To identify possible false positives due to mismapped reads, we implemented the following filtering steps: As suggested in prior discussion of this topic (Li, 2014) we used the coverage information at the variant sites to tag possible false positives. However, coverage information alone is too coarse for the resolution of low frequent FP. Therefore we additionally monitor break positions that flank or reside at the variant positions to identify regions with mismapped reads. As all FP were present as small clusters or hotspots, we tagged regions that hosted more than 4 SNPs within a sliding window at the double length of the insert size and were accompanied by a break position (BP) as possible FP causing regions. With the application of these filters we could identify and exclude the FP calls (**Fig. 7 C**).

A closer look at inversions revealed, that they were mostly not identified as inversions but the start and the end point of the inversion were marked as break positions (**Supp. Table 1**). Break positions occur because only one read of a pair can be mapped, leading to an accumulation of only forward or reverse reads. They indicate a larger sequence difference between the reads and the reference and are therefore a more general indicator of a larger structural variation. Therefore these calls represent a partial resolution of the variant.

3.2.3 Performance of combined post-processing and filtering in VarCap

We observed that a gain in variant calling sensitivity decreased the accuracy. Therefore we added a post-filtering step to the workflow in order to eliminate possible FP. We incorporated a post-processing step for each variant that aims to eliminate FP due to sequencing errors, repetitive regions, partially detected variants and mismapped reads due to reference incompleteness. As a consequence of the dissimilar variant detection rates of some methods, we decided to use more than one tool for each type of variant. In order to gain accuracy and robustness, for high confidence variants, a variant call had to be supported by at least two different tools. This step further contributed to an improved accuracy at low MRA cutoffs (1%), while the detection rate was only slightly diminished (**Table 1**).

3.3 Genotyping of diverse synthetic prokaryotic populations

3.3.1 Detection rates in different genomes

Genomes exhibit different properties, such as G+C content and size, which could potentially affect the sensitivity and accuracy of variant calling. Therefore we evaluated our variant calling workflow on six different genomes. These organisms consisted of five bacteria and one archaeon, with differing G+C content ranging from 26 to 72 percent as well as a differing genome size ranging from 0.68 to 8.66 Mb. The workflow was used with a MRA of 2% as well as at a MAA of 8 reads supporting a variation. In concordance to our previous results we could detect most of the (simulated) variants (>90%). However, at a MRA of 2% we could not observe any dependency on G+C content or genome size while the MAA of 8 reads resulted in fewer variant detections at high G+C content and genome size (**Fig. 8**). Therefore we decided to use the more flexible MRA as a minimum boundary for variant detection as it showed less influence to different genome properties.

3.3.2 Detection rates in a distantly evolved population

More distantly evolved populations carry higher variant frequencies as tested before, which could affect the sensitivity of variant calling. Therefore ALFSim (Dalquen et al. 2012) was used to simulate a more distantly evolved population by integrating evolutionary changes (SNPs, Indels and duplications) into the *P. amoebophila* genome. The evolved genome showed a similarity to the reference around 99%, as it contained around 21000 SNPs, 100 Indels and three gene duplications.

We evaluated the sensitivity of the variant calling by VarCap at a low abundant subpopulation of 4%. We used a MRA of 3%, 2% and 1% as well as a MAA of eight reads (equals a MRA of 2% in a 400x covered genome). Depending on the minimum abundance requirements, we were able to detect between 90% and 99% of all SNPs, between 74% and 94% of all indels and 2 out of 3 duplications. The true positive detection rate of SNPs increased to 98%, while the false positive rate remained below 0.3% when lowering the MRA from 3 to 2%. However, if we lowered MRA further to 1%, we increased the TP rate to 99% while augmenting the FP rate close to 400 FP/Mb (**Fig. 9A**). At a MRA of 2% we could locate most FP within repetitive regions and recent duplications (**Fig. 9B**), while at a MRA of 1% we detected mainly FP caused by the sequencing error rate (**Fig. 9B**). At a MRA of 2%, we were able to detect over 90% of all Indels including all small Indels (size=1), without experiencing false positives (**Fig. 9A**). With regard to duplications we were able to find two of them at most MRAs, while missing out the shortest one constantly (**Fig. 9C SV(DUP)**). These findings confirm that we are able to achieve a high accuracy even if the evolved genomes are rather dissimilar. However, a novel finding was that also recent duplications can lead to wrongly placed reads as they are similar to repetitive regions. Therefore we also included tagging of duplicated regions as possible regions for FP calls into our workflow.

3.4 Detecting variants in a real bacterial population after long term cultivation

In order to predict variant frequencies within an evolving population, the variant calling workflow was applied to a long-term cultivation experiment of *P. amoebophila*. Different MRA cutoffs from 20% to 2% were used and revealed that variants were present at frequencies down to 2% (**Fig. 10A**, outer rings). Variants within repetitive regions (**Fig. 10A**, inner connective lines) were tagged for further inspection. At a MRA of 2% we observed a total number of 71 variants, which comprised of 34 SNPs, 20 Indels and 17 structural variants. The SNPs and small Indels were annotated using SNPEff (Cingolani et al. 2012). This revealed, that around 83% of them were situated within coding regions (**Supp. Table 3**). At a MRA of 2% we could find three Indels present at a subpopulation of 2%. They are located within homopolymeric regions of length 10 (data not shown) and thus were tagged as probable FP for further manual inspection.

For the validation of the variant calling prototype of VarCap we picked three variations for further analysis that were present at abundances of 4%,11% and 28%, accordingly. We performed PCR of the regions surrounding the three variants, cloned the fragments into vectors and picked 16 clones of each variant for Sanger sequencing (**Table 2**, **Supp. Table 4**). We were able to detect all three variants and thus could confirm the predictions of the VarCap software.

4. DISCUSSION

Population genomics of microbes is most powerful if we meet the challenge to detect all types of genomic variations even at low frequency. We therefore developed, evaluated and validated VarCap, a workflow that allowed us to reliably identify variants even within low abundant alleles. We tested the capabilities of the relevant variant calling tools and observed substantial sensitivity differences between the different methods. In order to improve the overall sensitivity we decided to integrate different tools for variant detection into a combined workflow, in which every variant can be detected by more than one caller. As more tools are likely to introduce more errors, we also optimized the overall accuracy. Detecting sequencing errors and mis-mapped reads was key to control the rate of false positives. We observed that for SNPs a MRA of 2% was sufficient (MAA of 8 reads) to keep a safety margin to false positives appearing at a MRA of 1%. This implies, that for detecting a subpopulation present at 2% we need a minimum sequencing coverage of 400x. Sequencing experiments should therefore aim for at least 500x to account for reads removed by quality filtering and fluctuations in coverage along the genome. We could not detect any FP Indels within our simulated data but detected several spurious Indels in homopolymer regions of the re-sequencing experiment. These are probably sequencing/PCR artifacts that are not introduced by read simulators. Based on our findings Indels below a MRA of 10% should be tagged as potentially false positive if they are located within a homopolymeric region. Mismapped reads can occur within repetitive regions, undetected duplications, or incomplete reference genomes. Therefore we flag repetitive regions greater than the insert size in order to mark variants appearing within these regions for further inspection. Unnoticed duplications or incomplete references cause reads to get mapped to similar regions, which can be observed by higher coverage and/or variant accumulation within these areas. In order to overcome false positives by misplaced reads, we suggest tagging variants that at least fulfill two of the three following rules: I) Either variants lie within regions with a coverage of 20% above the average and/or II) if there is a break position detected at or within read length of the variant site and/or III) if they lie within a repeat region and/or IV) if more than 5 variants lie within the length of one insert size. Furthermore, for extracting high confidence variants, we requested each variant to be confirmed by at least two callers.

We observed that insertions and especially inversions were harder to detect than the rest of the variations. This is not unexpected, as current methods for their prediction need sufficient support by reads, which may get lost at low frequencies. In the simulated evolution data we missed the shortest duplication constantly. This may be related to a combination of callers working at their operational limits (300 bp) and a diverging evolution of the duplicated sequence due to newly introduced SNPs. According to our results, we could establish rules for filtering out errors and help with the interpretation of different types of variations (e.g. SNP, duplications). Using these rules we have built a fully automated workflow that reliably predicts rare variants in deep sequencing data.

5. CONCLUSION

We created VarCap, a fully automated workflow that allows scientists to rapidly predict variants within high coverage, short read paired end sequencing data. VarCap automatically performs quality filtering, mapping, variant calling and post-filtering of the predicted variants. In order to allow a broad community to use VarCap, we implemented VarCap within our Galaxy Server, which is publicly available at http://galaxy.csb.univie.ac.at. VarCap includes default parameter settings, derived from our evaluation experiments, to keep it for the user as simple as possible. The output of VarCap is a vcf file with a detailed description of the variants and two pdf files, which give a graphical overview of variant coverage and their frequency distribution. VarCap is designed to predict different allele frequencies in experimental evolution experiments, and it is able to detect and report the frequencies of multiple genotypes within clinical samples e.g. multiple infections.

FUNDING

This work was funded by a grant from the European Research Council (ERC StG EVOCHLMY, grant no. 281633).

REFERENCES

Afgan, Enis, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, et al. 2016. “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update.” *Nucleic Acids Research*, May, gkw343. doi:10.1093/nar/gkw343.

Barrick, Jeffrey E, and Richard E Lenski. 2013. “Genome Dynamics during Experimental Evolution.” *Nature Reviews. Genetics* 14 (12): 827–39. doi:10.1038/nrg3564.

Blount, Zachary D., Jeffrey E. Barrick, Carla J. Davidson, and Richard E. Lenski. 2012. “Genomic Analysis of a Key Innovation in an Experimental Escherichia Coli Population.” *Nature* 489 (7417): 513–18. doi:10.1038/nature11514.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics (Oxford, England)* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.

Bos, Kirsten I., Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, et al. 2011. “A Draft Genome of Yersinia Pestis from Victims of the Black Death.” *Nature* 478 (7370): 506–10. doi:10.1038/nature10549.

Chen, Jian-Qun, Ying Wu, Haiwang Yang, Joy Bergelson, Martin Kreitman, and Dacheng Tian. 2009. “Variation in the Ratio of Nucleotide Substitution and Indel Rates across Genomes in Mammals and Bacteria.” *Molecular Biology and Evolution* 26 (7): 1523–31. doi:10.1093/molbev/msp063.

Chen, Ken, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath, et al. 2009. “BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6 (9): 677–81. doi:10.1038/nmeth.1363.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3.” *Fly* 6 (2): 80–92. doi:10.4161/fly.19695.

Dalquen, Daniel A., Maria Anisimova, Gaston H. Gonnet, and Christophe Dessimoz. 2012. “ALF—A Simulation Framework for Genome Evolution.” *Molecular Biology and Evolution* 29 (4): 1115–23. doi:10.1093/molbev/msr268.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98. doi:10.1038/ng.806.

Earl, Dent, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. “Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods.” *Genome Research* 21 (12): 2224–41. doi:10.1101/gr.126599.111.

Eyre, David W., Madeleine L. Cule, David Griffiths, Derrick W. Crook, Tim E. A. Peto, A. Sarah Walker, and Daniel J. Wilson. 2013. “Detection of Mixed Infection from Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in Clostridium Difficile Transmission.” *PLOS Comput Biol* 9 (5): e1003059. doi:10.1371/journal.pcbi.1003059.

Gardy, Jennifer L., James C. Johnston, Shannan J. Ho Sui, Victoria J. Cook, Lena Shah, Elizabeth Brodkin, Shirley Rempel, et al. 2011. “Whole-Genome Sequencing and

Social-Network Analysis of a Tuberculosis Outbreak.” *New England Journal of Medicine* 364 (8): 730–39. doi:10.1056/NEJMoa1003176.

Horn, Matthias, Astrid Collingro, Stephan Schmitz-Esser, Cora L Beier, Ulrike Purkhold, Berthold Fartmann, Petra Brandt, et al. 2004. “Illuminating the Evolutionary History of Chlamydiae.” *Science (New York, N.Y.)* 304 (5671): 728–30. doi:10.1126/science.1096330.

Hu, Xuesong, Jianying Yuan, Yujian Shi, Jianliang Lu, Binghang Liu, Zhenyu Li, Yanxiang Chen, et al. 2012. “pIRS: Profile-Based Illumina Pair-End Reads Simulator.” *Bioinformatics (Oxford, England)* 28 (11): 1533–35. doi:10.1093/bioinformatics/bts187.

Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. 2012. “De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs.” *Nature Genetics* 44 (2): 226–32. doi:10.1038/ng.1028.

Khan, Aisha I., Duy M. Dinh, Dominique Schneider, Richard E. Lenski, and Tim F. Cooper. 2011. “Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population.” *Science* 332 (6034): 1193–96. doi:10.1126/science.1203801.

Koboldt, Daniel C, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. 2012. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research* 22 (3): 568–76. doi:10.1101/gr.129684.111.

Kofler, Robert, and Christian Schlötterer. 2014. “A Guide for the Design of Evolve and Resequencing Studies.” *Molecular Biology and Evolution* 31 (2): 474–83. doi:10.1093/molbev/mst221.

Koonin, Eugene V., and Yuri I. Wolf. 2008. “Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World.” *Nucleic Acids Research* 36 (21): 6688–6719. doi:10.1093/nar/gkn668.

Köser, Claudio U., Matthew T.G. Holden, Matthew J. Ellington, Edward J.P. Cartwright, Nicholas M. Brown, Amanda L. Ogilvy-Stuart, Li Yang Hsu, et al. 2012. “Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak.” *New England Journal of Medicine* 366 (24): 2267–75. doi:10.1056/NEJMoa1109910.

Li, Heng. 2014. “Toward Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.” *Bioinformatics* 30 (20): 2843–51. doi:10.1093/bioinformatics/btu356.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Maharjan, Ram P., Joël Gaffé, Jessica Plucain, Martin Schliep, Lei Wang, Lu Feng, Olivier Tenaillon, Thomas Ferenci, and Dominique Schneider. 2013. “A Case of Adaptation through a Mutation in a Tandem Duplication during Experimental Evolution in Escherichia Coli.” *BMC Genomics* 14 (1): 441. doi:10.1186/1471-2164-14-441.

McElroy, Kerensa, Torsten Thomas, and Fabio Luciani. 2014. “Deep Sequencing of Evolving Pathogen Populations: Applications, Errors, and Bioinformatic Solutions.” *Microbial Informatics and Experimentation* 4 (1): 1. doi:10.1186/2042-5783-4-1.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. doi:10.1101/gr.107524.110.

Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. “Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems.” *Genome Biology* 12 (11): R112. doi:10.1186/gb-2011-12-11-r112.

Moran, Nancy A., Heather J. McLaughlin, and Rotem Sorek. 2009. “The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria.” *Science* 323 (5912): 379–82. doi:10.1126/science.1167140.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. “Genotype and SNP Calling from next-Generation Sequencing Data.” *Nature Reviews Genetics* 12 (6): 443–51. doi:10.1038/nrg2986.

Patel, Ravi K., and Mukesh Jain. 2012. “NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data.” *PLoS ONE* 7 (2): e30619. doi:10.1371/journal.pone.0030619.

Pruitt, Kim D., Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. 2012. “NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy.” *Nucleic Acids Research* 40 (D1): D130–35. doi:10.1093/nar/gkr1079.

Pulido-Tamayo, Sergio, Aminael Sánchez-Rodríguez, Toon Swings, Bram Van den Bergh, Akanksha Dubey, Hans Steenackers, Jan Michiels, Jan Fostier, and Kathleen Marchal. 2015. “Frequency-Based Haplotype Reconstruction from Deep Sequencing Data of Bacterial Populations.” *Nucleic Acids Research*, May, gkv478. doi:10.1093/nar/gkv478.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65. doi:10.1038/nature08821.

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. 2012. “DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis.” *Bioinformatics* 28 (18): i333–39. doi:10.1093/bioinformatics/bts378.

Schmieder, Robert, and Robert Edwards. 2011. “Quality Control and Preprocessing of Metagenomic Datasets.” *Bioinformatics*, January, btr026. doi:10.1093/bioinformatics/btr026.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. 2012. “LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets.” *Nucleic Acids Research* 40 (22): 11189–201. doi:10.1093/nar/gks918.

Yang, Xiao, Sriram P. Chockalingam, and Srinivas Aluru. 2013. “A Survey of Error-Correction Methods for next-Generation Sequencing.” *Briefings in Bioinformatics* 14 (1): 56–66. doi:10.1093/bib/bbs015.

Ye, Kai, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009. “Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads.” *Bioinformatics (Oxford, England)* 25 (21): 2865–71. doi:10.1093/bioinformatics/btp394.

Yost, Shawn E., Hakan Alakus, Hiroko Matsui, Richard B. Schwab, Kristen Jepsen, Kelly A. Frazer, and Olivier Harismendy. 2013. “Mutascope: Sensitive Detection of Somatic Mutations from Deep Amplicon Sequencing.” *Bioinformatics*, May, btt305. doi:10.1093/bioinformatics/btt305., May, btt305. doi:10.1093/bioinformatics/btt305.

Table 1

Min Caller	MRA	TP	FN	FP	Sensitivity	Precision
1	10	139	2	0	0.986	1
	5	137	4	0	0.972	1
	2	138	3	0	0.979	1
	1	141	0	1238	1.000	0.102
2	10	135	6	0	0.957	1
	5	133	8	0	0.943	1
	2	133	8	0	0.943	1
	1	135	6	0	0.957	1

Table 1 Analysis of the detection sensitivity of the combined workflow for different number of callers and various MRAs. The table shows the numbers for the observed true positives (TP), false negatives (FN), false positives (FP), sensitivity and precision of the combined workflow at MRAs of 10, 5, 2 and 1% under the requirement that either 1 or 2 callers (Min Caller) had to confirm each variant.

Table 2

Position	Frequency	Clones total	Clones supportive	Sanger confirmed
1338568	28	16	6	Yes
1339720	11	16	2	Yes
1339224	4	16	1	Yes

Table 2 Experimental validation of a subset of the predicted variations. Three variant positions at different frequencies were amplified by PCR, cloned and Sanger sequenced for validation.

Figure 1

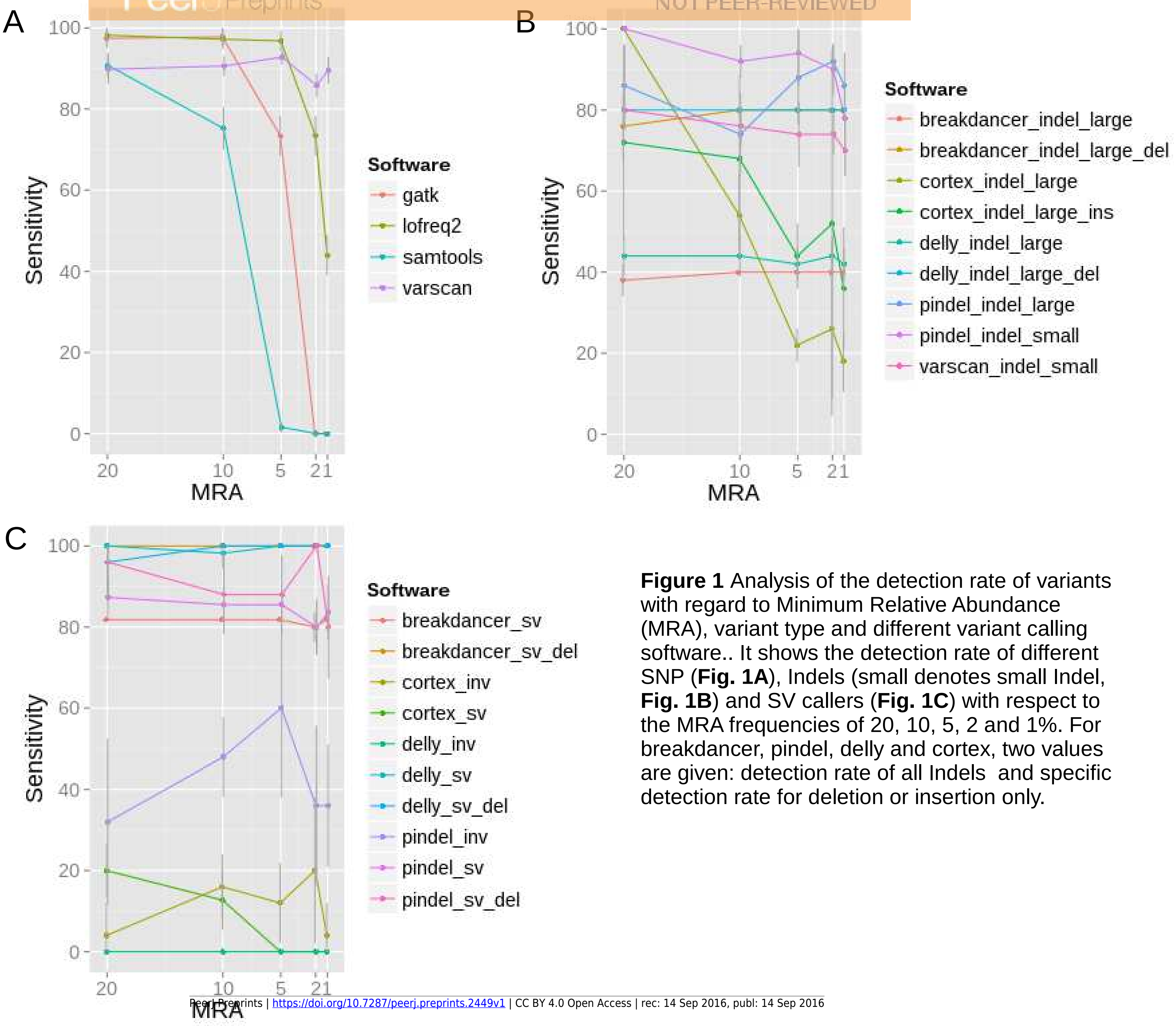
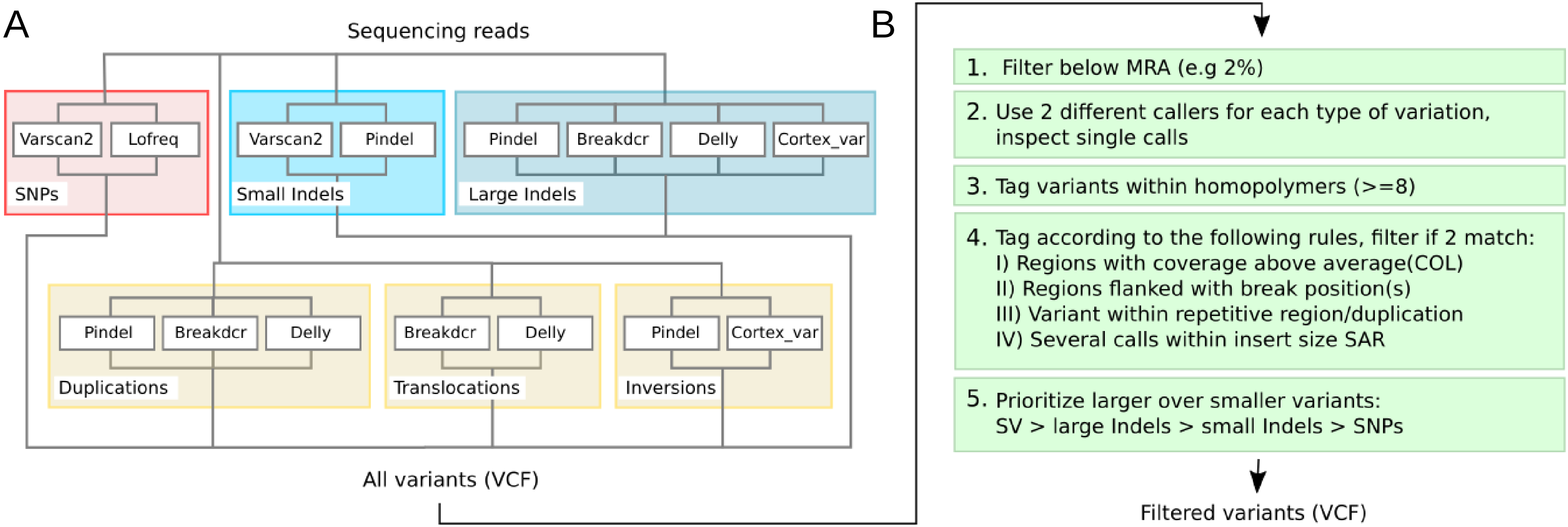


Figure 1 Analysis of the detection rate of variants with regard to Minimum Relative Abundance (MRA), variant type and different variant calling software.. It shows the detection rate of different SNP (**Fig. 1A**), Indels (small denotes small Indel, **Fig. 1B**) and SV callers (**Fig. 1C**) with respect to the MRA frequencies of 20, 10, 5, 2 and 1%. For breakdancer, pindel, delly and cortex, two values are given: detection rate of all Indels and specific detection rate for deletion or insertion only.

Figure 2



1. Filter below MRA (e.g 2%)

2. Use 2 different callers for each type of variation, inspect single calls

3. Tag variants within homopolymers (>=8)

4. Tag according to the following rules, filter if 2 match:
I) Regions with coverage above average(COL)
II) Regions flanked with break position(s)
III) Variant within repetitive region/duplication
IV) Several calls within insert size SAR

5. Prioritize larger over smaller variants:
SV > large Indels > small Indels > SNPs

Figure 2 Detection capabilities of different tools at low frequencies. **Fig. 2A** shows the variant types that were successfully detected by the different software tools while **Fig. 2B** shows the postfiltering steps to eliminate false positives.

Figure 3

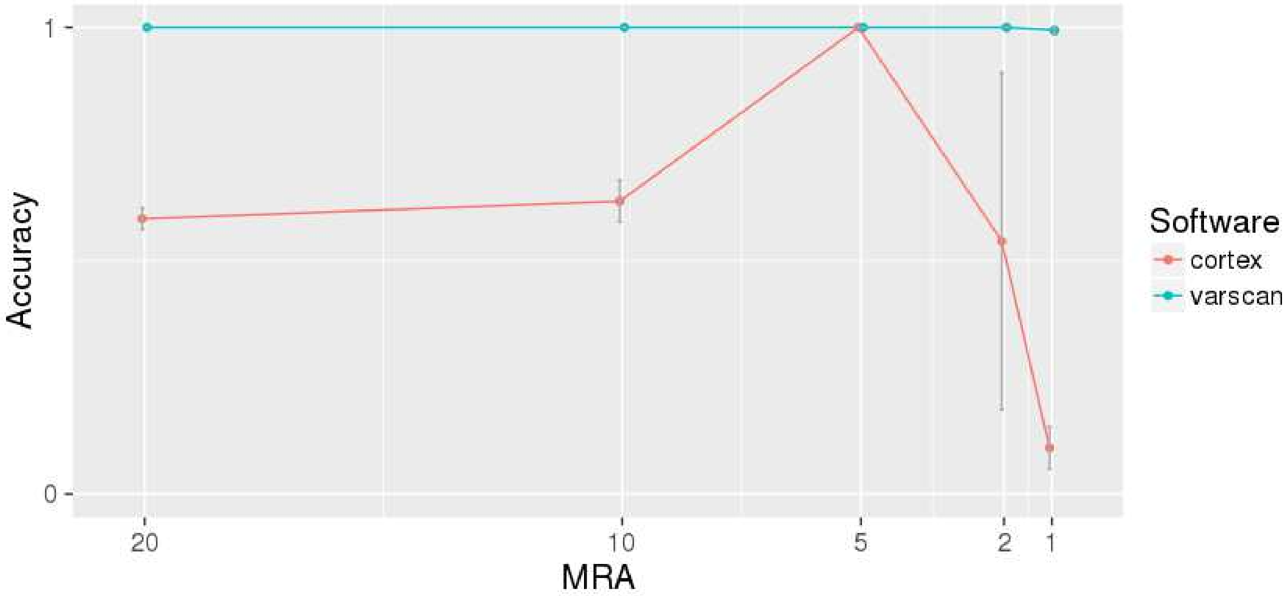


Figure3 Accuracy of the different variant calling tools. We visualized only Cortex_var and Varscan2 as they were the only tools that produced false positives after applying the MRA filters.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Figure 4

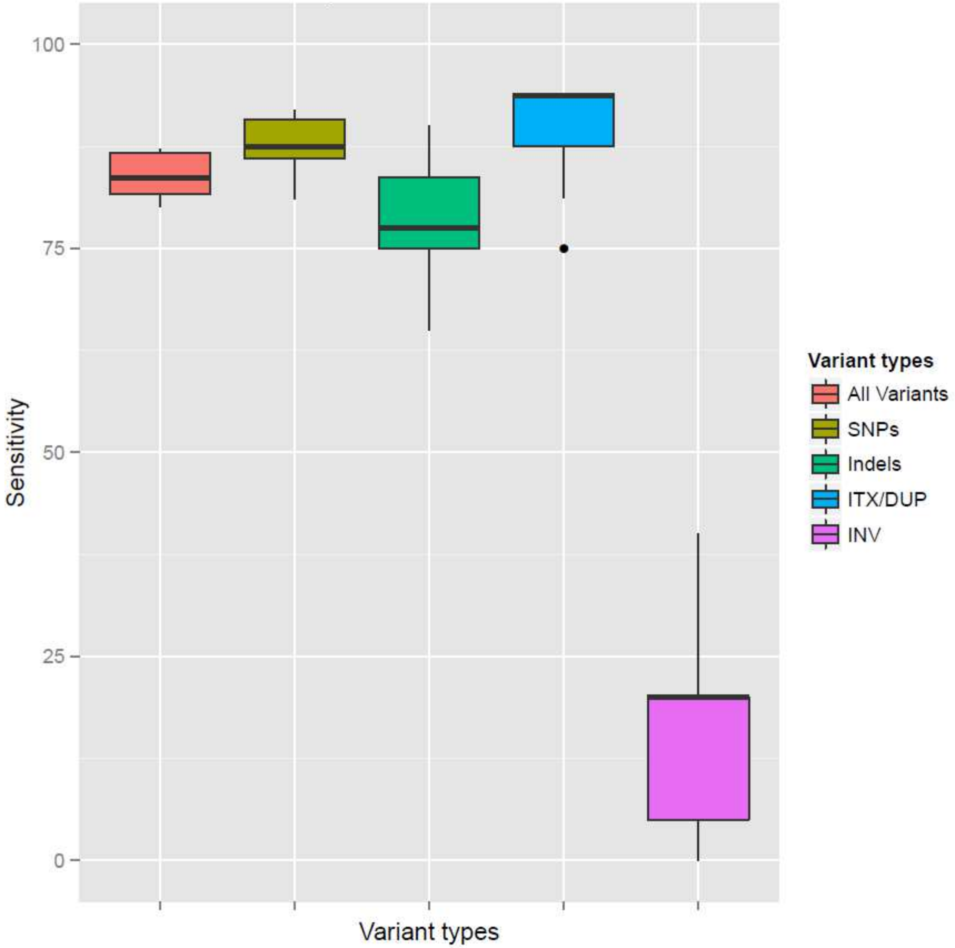


Figure 4 Average detection rates and standard deviation of all callers for variants >= MRA of 2%. The variant detection rates are shown in percent for all variants(ALL), only SNPs(SNP), only Indels (IND), duplications and translocations (ITX/DUP) and inversions (INV).

Figure 5

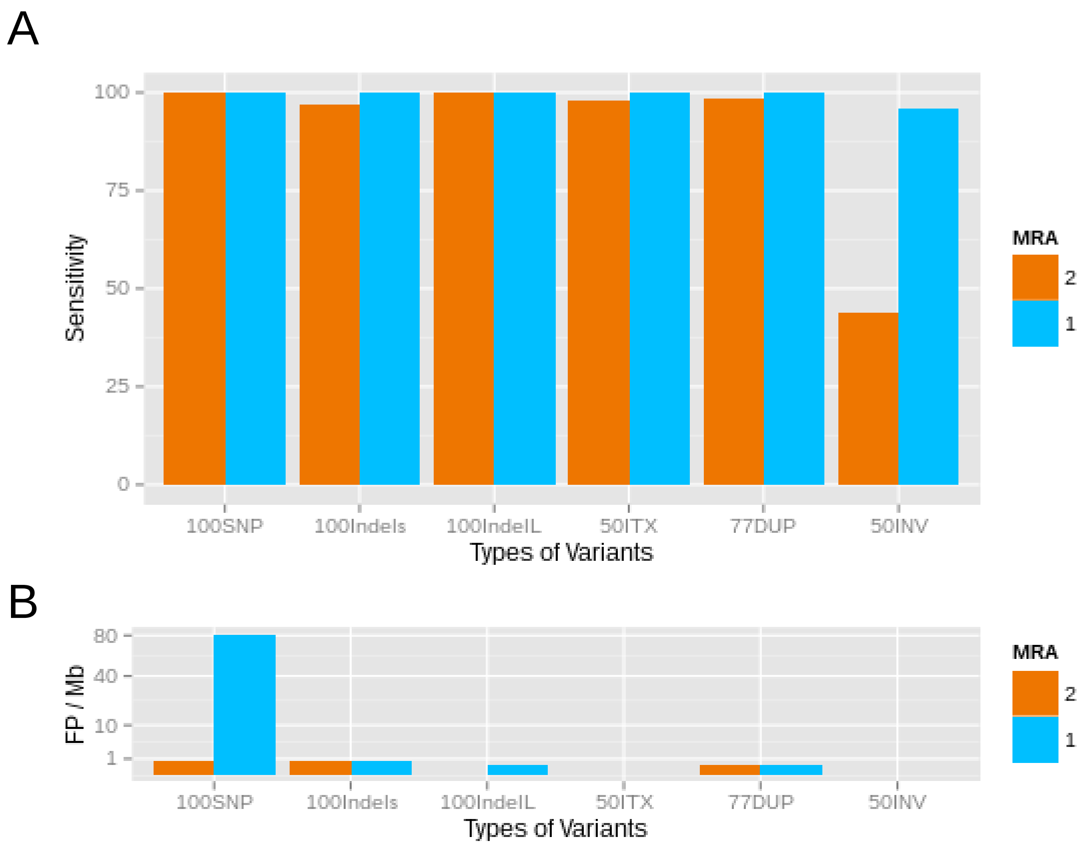
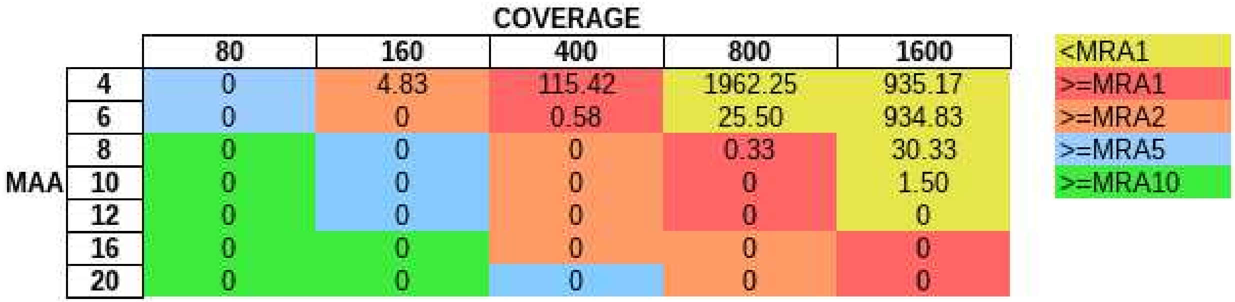


Figure 5 Variant detection rate of subpopulations that harbour one type of variant exclusively. **5A** shows the detection rate of True Positives in % at a MRA of 2% and MRA of 1% as well as False Positives per Megabase (MRA 2% , MRA 1% **5B**) according to different types of variants which were inserted separately into the *P. amoebophila* reference genome. The total coverage is at 400x, the coverage of the subpopulations containing either 100 or 77 or 50 variants is at 16x.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Figure 6



		COVERAGE				
		80	160	400	800	1600
MAA	4	🟦 0	🟧 4.83	🟥 115.42	🟨 1962.25	🟨 935.17
	6	🟦 0	🟧 0	🟥 0.58	🟨 25.50	🟨 934.83
	8	🟩 0	🟦 0	🟧 0	🟥 0.33	🟨 30.33
	10	🟩 0	🟦 0	🟧 0	🟥 0	🟨 1.50
	12	🟩 0	🟦 0	🟧 0	🟥 0	🟨 0
	16	🟩 0	🟩 0	🟧 0	🟧 0	🟥 0
	20	🟩 0	🟩 0	🟦 0	🟧 0	🟥 0

🟨	<MRA1
🟥	>=MRA1
🟧	>=MRA2
🟦	>=MRA5
🟩	>=MRA10

Figure 6 Influence of total coverage and MAA on FP rate. The table numbers show the FP per Megabase in context to coverage and MAA while the different colours indicate the corresponding MRA levels.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Figure 7

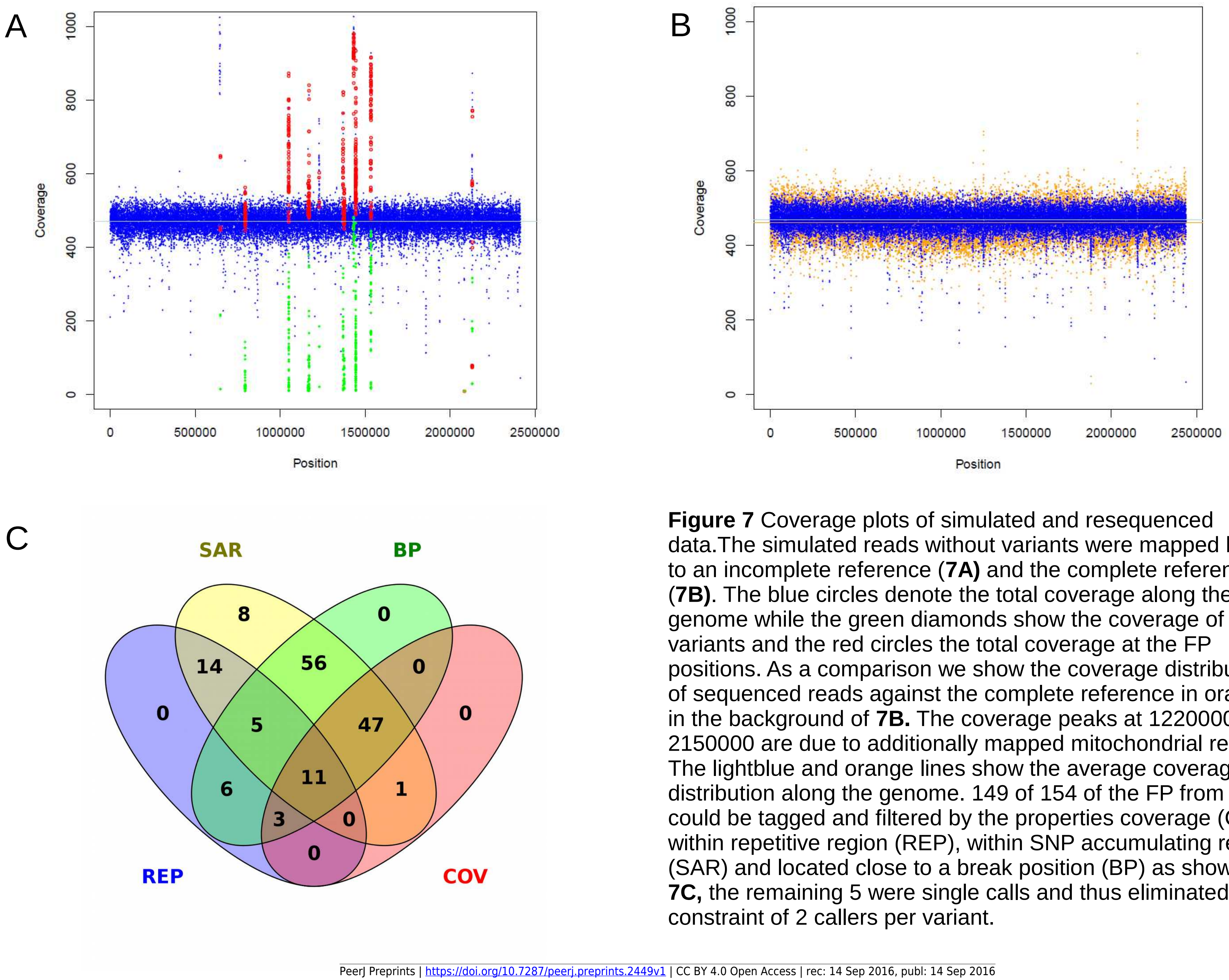


Figure 7 Coverage plots of simulated and resequenced data.The simulated reads without variants were mapped back to an incomplete reference (**7A)** and the complete reference (**7B)**. The blue circles denote the total coverage along the genome while the green diamonds show the coverage of the FP variants and the red circles the total coverage at the FP positions. As a comparison we show the coverage distribution of sequenced reads against the complete reference in orange in the background of **7B.** The coverage peaks at 1220000 and 2150000 are due to additionally mapped mitochondrial reads. The lightblue and orange lines show the average coverage distribution along the genome. 149 of 154 of the FP from **7A** could be tagged and filtered by the properties coverage (COV), within repetitive region (REP), within SNP accumulating region (SAR) and located close to a break position (BP) as shown in **7C,** the remaining 5 were single calls and thus eliminated by the constraint of 2 callers per variant.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Figure 8

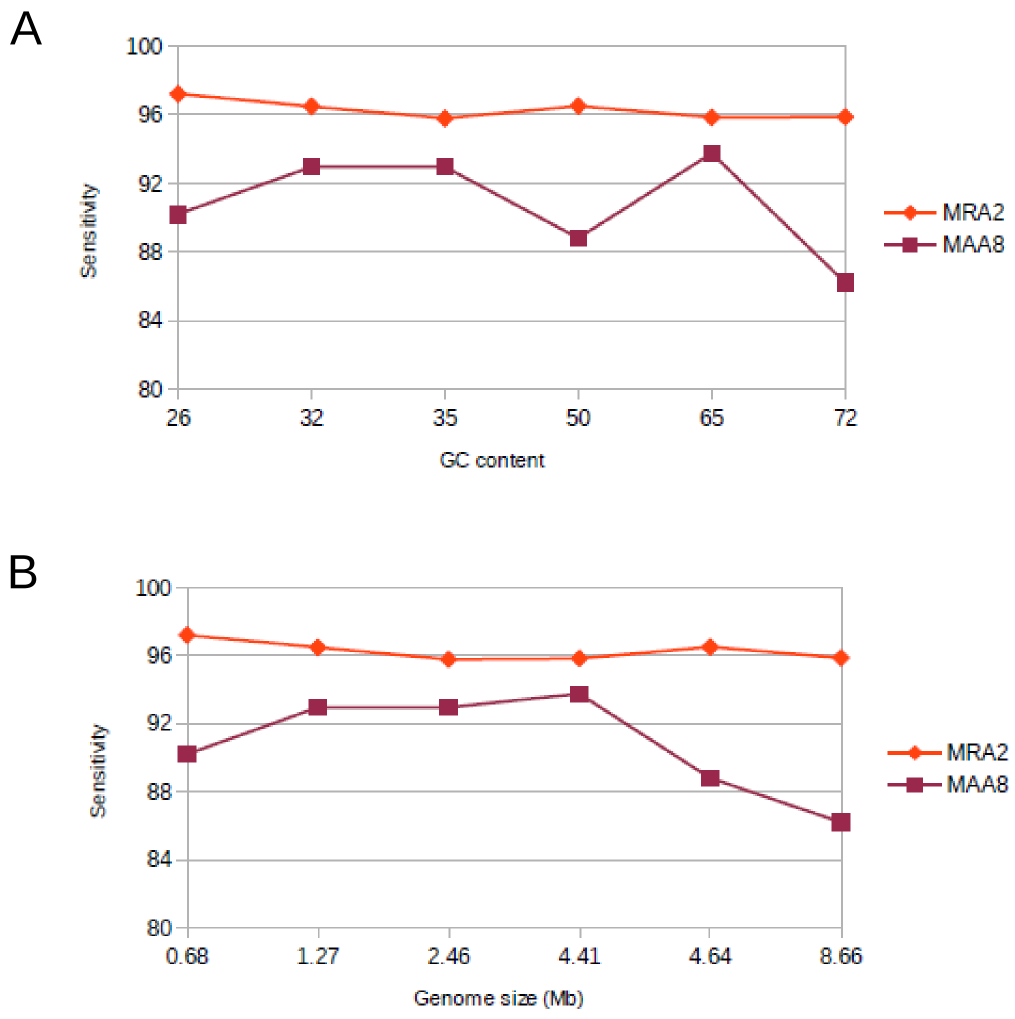


Figure 8 Detection rate of variants in various genomes at minimum absolute and relative abundence. The observed percentage of True Positives is shown for six organisms with differing GC content (**8A**) and genome size (**8B**). The total coverage is at 400x, the coverage of the subpopulation containing 135 variants is at 16x. No False Positives were observed at the MAA of 8 and MRA of 2%.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Figure 9

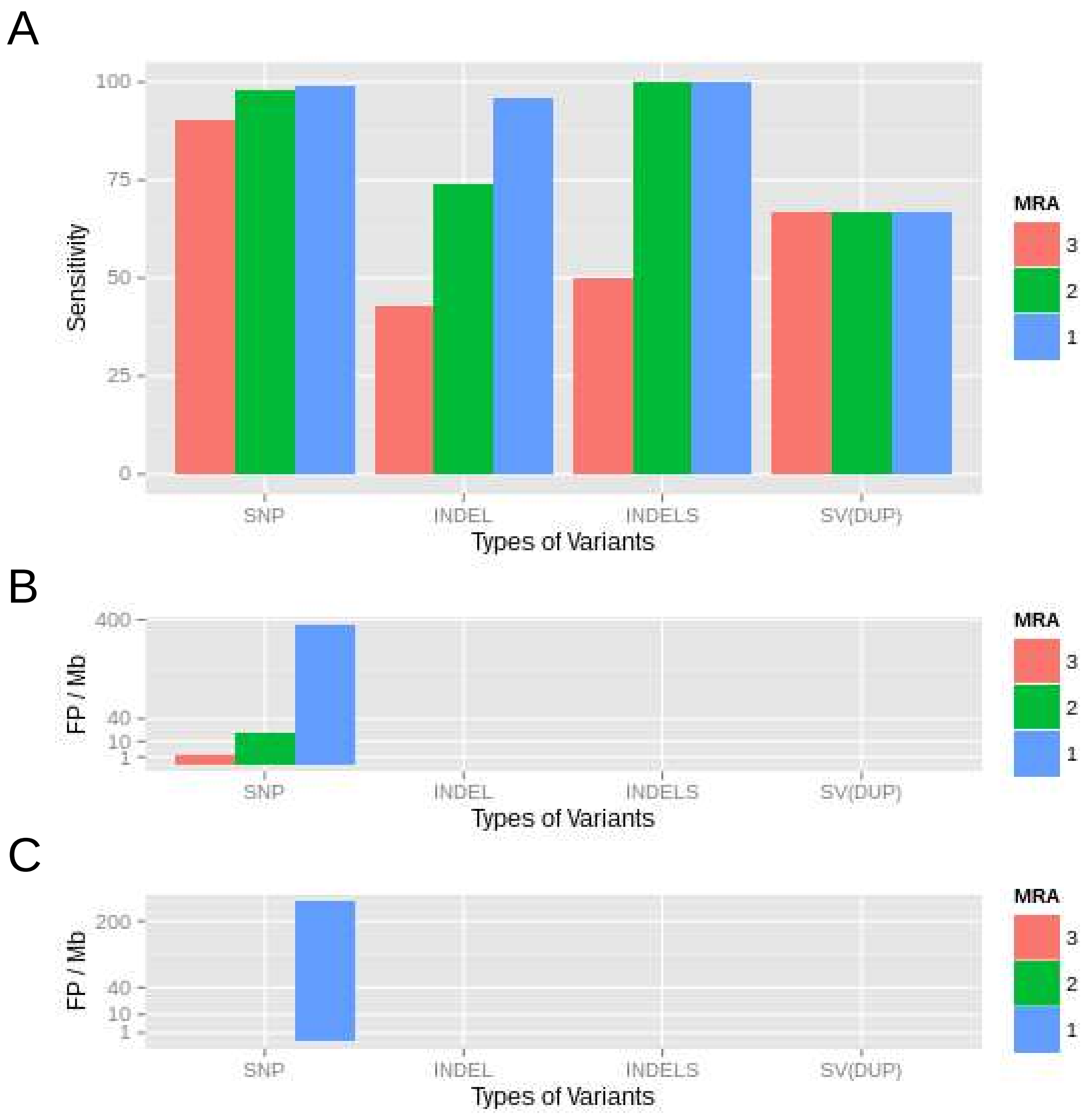
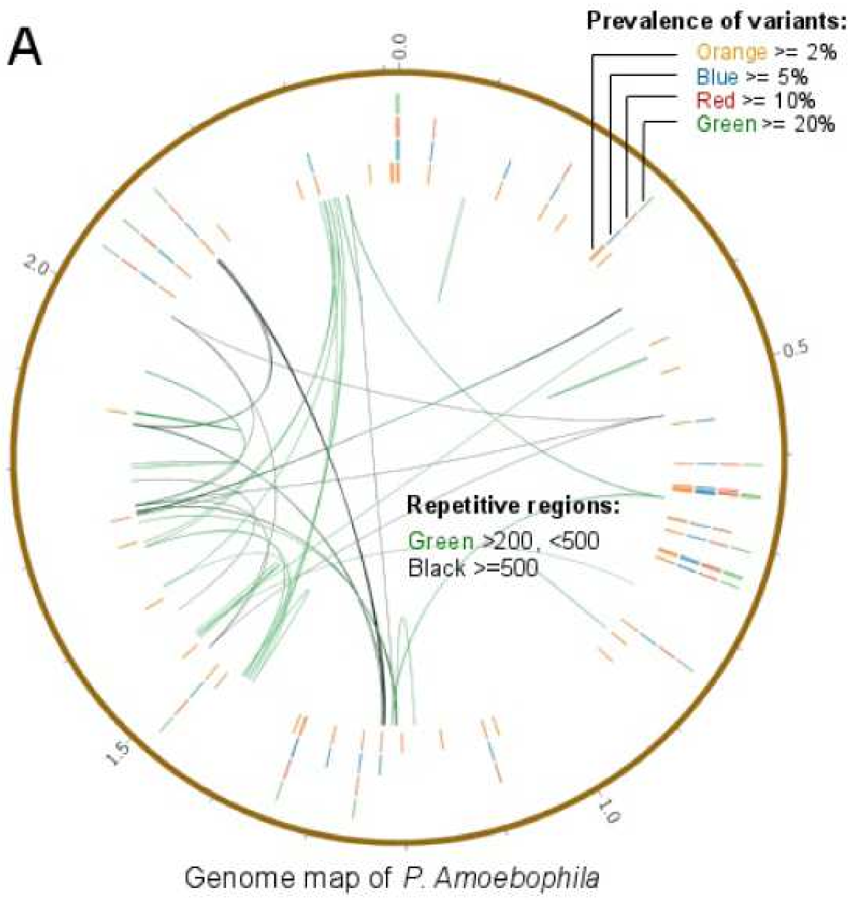


Figure 9 Observed detection rates of variants which were simulated using a genome evolution software (ALFSim) and detected at different minimum abundances. **9A** shows the sensitivity at MRAs of 3,2 and 1%. **9B** shows the False Positives for SNPs as counts per Magabase at the different MRAs. At these minimum abundances no FP for Indels and SV were detected. **9C** shows the FP per Megabase after filters have been applied.

Figure 10



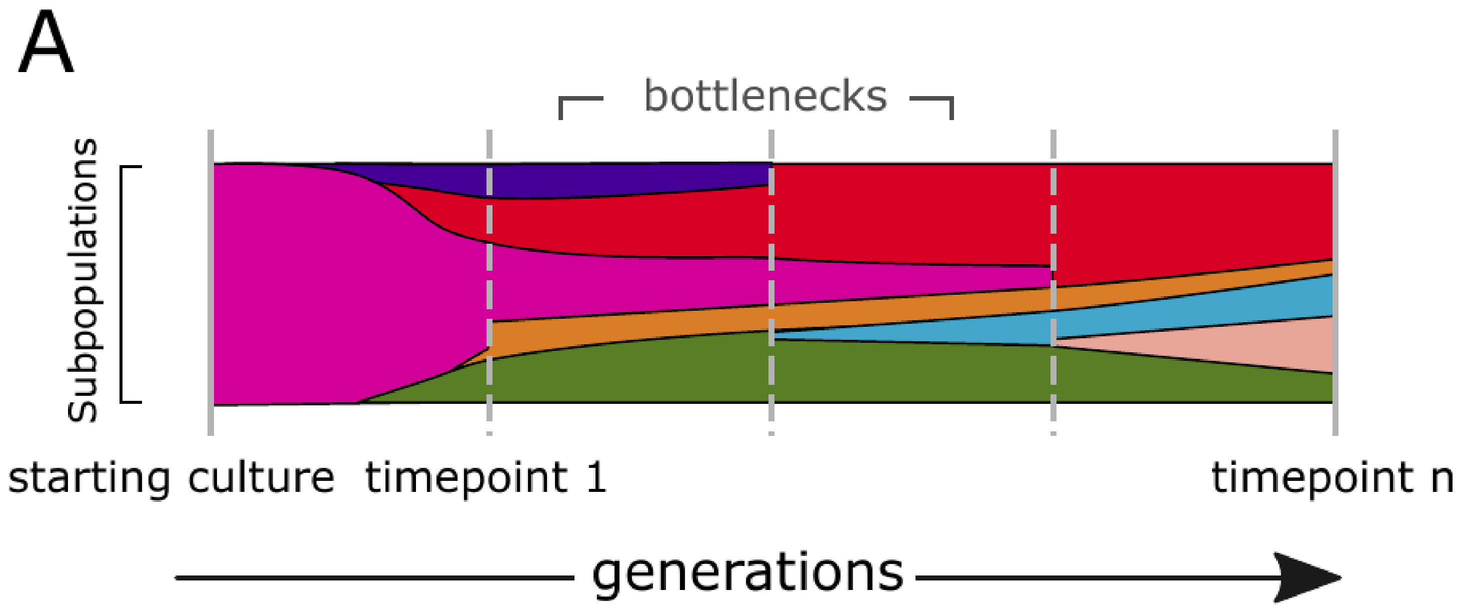
B

	MRA 2	MRA 5	MRA 10	MRA 20
SNP	34	23	16	11
Indel	20	8	8	5
SV	17	9	8	8

Figure 10 Prevalence of variants within a long term culture with respect to their MRAs. **10A** shows the prevalence of variations at MRAs of 20%, 10%, 5% or 2%, which are visible in the four differently colored outer circles and the presence of repetitive regions within the reference genome (inner connective lines). **10B** shows a more detailed view of variations found at MRAs of 20%, 10%, 5% and 2%.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016

Introduction figures



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2449v1 | CC BY 4.0 Open Access | rec: 14 Sep 2016, publ: 14 Sep 2016