WAGENINGEN
UNIVERSITY & RESEARCH

SNP detection in allopolyploid crops using NGS data

Abstract

Homologous SNP detection in polyploid organisms is complicated due to the presence of subgenome polymorphisms, i.e. homeologous SNPs. Several filtering tools have been developed to distinguish between homologous SNPs and homeologous SNPs. We have studied one of these filtering tools, SWEEP, using simulated NGS data. Here we show that SWEEP is not the optimal choice when dealing with allopolyploids with high ploidy level. We modified SWEEP to improve its performance and developed a new approach based on FreeBayes. We show that the performance of SWEEP is limited by prior SNP calling and that FreeBayes returns almost all homologous SNPs at high ploidy levels.

María de la O Ferreras Gutiérrez
Supervisors: Ehsan Motazedi & Dick de Ridder

Introduction

Molecular breeding, known as the use of genetic manipulation to select and enhance a certain trait of interest, is widely used in present plant and animal breeding. A common tool applied by breeders is the use of genetic markers (Jiang, G.L. 2015).

Genetic markers are frequent alterations in DNA sequence, which allows to identify individuals from a population and track the inheritance through the generations (Encyclopaedia Britannica, 2017). Single nucleotide polymorphisms (SNPs) are the most popular genetic markers, due to their abundance in the genome and because the ease of detection, (Shaw, G. 2013; Kwok, P.Y. et al, 1998).

Several technologies for SNP detection have been developed, such as denaturing gradient gel electrophoresis (DGGE), SNP arrays or direct DNA sequencing. Years back, direct DNA sequencing was expensive and arduous, however, nowadays, next generation sequencing (NGS), which is faster and cheaper than previous sequencing methods, has allowed SNP detection through direct DNA sequencing (Kwok, P.Y. et al, 2003)

Even though SNP detection works well for diploid organisms, in polyploid organisms SNP detection is more challenging, due to the presence of polymorphisms between subgenomes.

An organism is called a polyploid when it contains more than two chromosome sets, i.e.

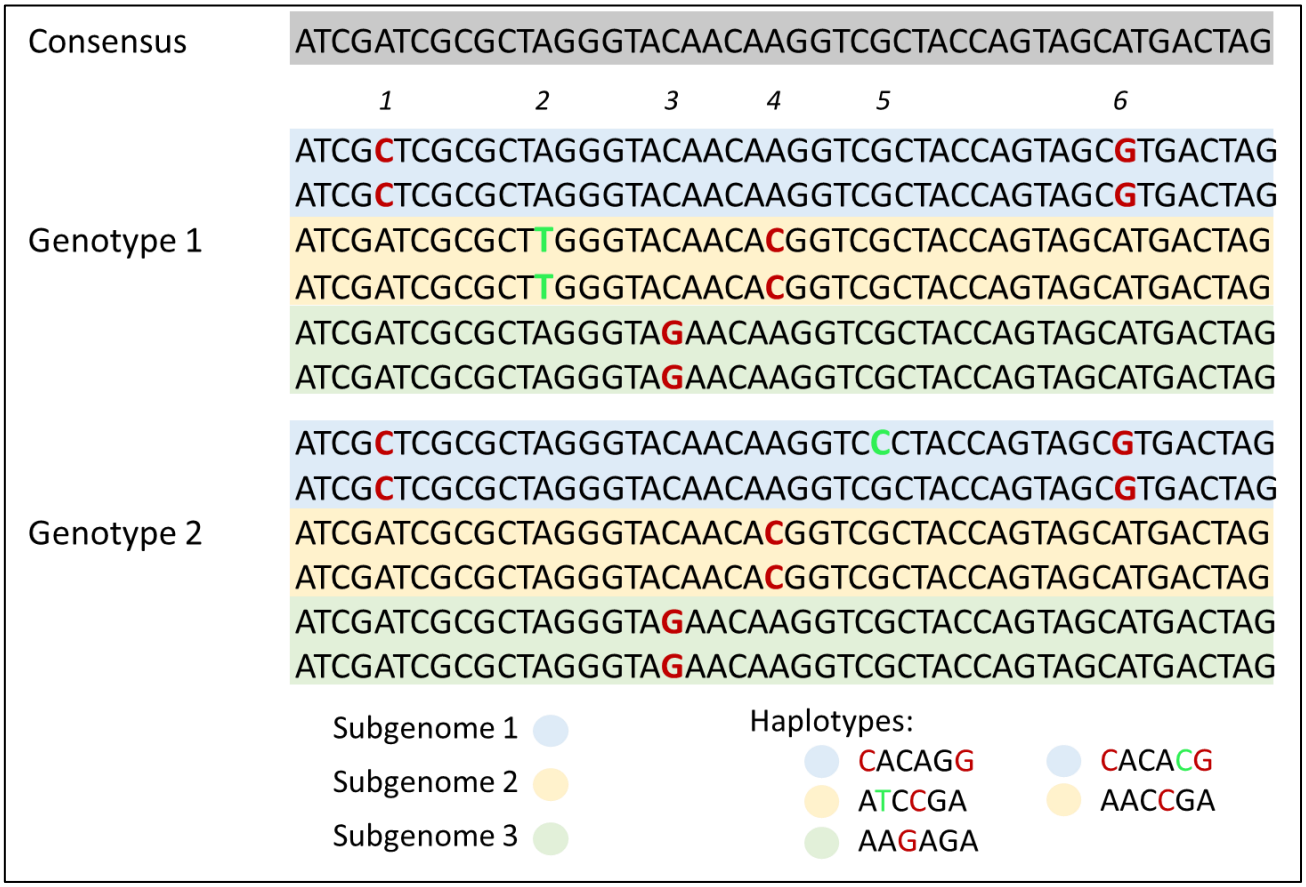


Figure 1. Two hexaploid organisms, i.e. genotypes, containing two different subgenomes presenting both homeologous SNPs (1,3, 4 and 6; marked in red) and homologous SNPs (2 and 5; marked in green).

group of chromosomes that carry the basic set of genetic information. Polyploids are divided in: autopolyploids, which contain several chromosome sets from only one species; and allopolyploids, which are hybrids from more than one related species, and therefore, contain two or more distinct genomes (Griffiths, A.J.F. et al, 2000).

In allopolyploids, due to the co-existence of different genomes in the same organism, copies within the same chromosome set are not identical, they are divided into subgenomes, depending on their input genome. Subgenomes are not completely homologous because they contain specific sequences of the genome of origin (see Figure 1) (Hirakawa, H. et al, 2014; Clevenger, J. et al, 2015)

There are two types of SNPs in allopolyploid organisms: homeologous SNPs, polymorphisms that represent differences between hybridized subgenomes; and homologous SNPs, polymorphisms that occur within the same subgenome (see Figure 1) (Clevenger, J. et al, 2015).

Within the set of homologous SNPs, we can distinguish two different types: heterozygous SNPs, which are not present in all the chromosome copies that belong to the same subgenome; and homozygous SNPs, which, in contrast, are present in all the chromosome copies that belong to the same subgenome. Examples of both types are given in Figure 1: SNP number 2 is an homozygous SNP, where both copies of subgenome 2 in genotype 1 express the alternate allele, T, while in genotype 2 all copies express the reference allele, A; and SNP number 5 is a heterozygous SNP, where one copy of subgenome 1 in genotype 2 expresses the alternate allele, C, while the other copy of genotype 2 and all copies of genotype 1 express the reference allele, G.

The reason why homologous SNPs are of interest lies in the haplotypes, defined as the group of SNPs that tend to be inherited together (Altshuler, D. et al, 2005). Figure 1 shows five different haplotypes: two for subgenome 1, two for subgenome 2 and one for subgenome 3. Subgenome 3 only provides one haplotype because there are no homologous SNPs present in that subgenome, only one homeologous SNP (SNP number 3), that is present in both genotypes and therefore if an individual inherits this haplotype, it is not possible to verify from which parent genotype it was inherited. However, for individuals who inherit the haplotypes containing homologous SNPs (CACACG and ATCCGA), these would be unique for the corresponding genotype and, therefore, it can be tracked.

SNP calling methods, which use “Samtools mpileup” in combination with “Bcftools call”, FreeBayes or GATK; these tools identify SNPs in the reads, but they also call homeologous SNPs, there is a need of an extra step to filter out homeologous SNPs.

The problem of distinguishing between homeologous SNPs and homologous SNPs has been addressed by different filtering methods, that are applied after SNP calling to eliminate homeologous SNPs, such as SWEEP (Sliding Window Extraction of Explicit

Polymorphisms), which considerably reduces the amount of homeologous SNPs in comparison with the traditional methods. (Clevenger, J. et al, 2015 a; Clevenger J. et al, 2015 b).

In this project we tested SWEEP in simulated NGS allopolyploid data in three different ploidy levels to identify weaknesses on its performance. We also assessed SWEEP and provided a pipeline that uses FreeBayes to perform SNP detection and two Python pipelines which perform the filtering step.

Materials & Methods

This section presents two pipelines used to simulate allopolyploid NGS data from a given reference genome. The simulated data processing and the tools used are also provided here. The last two subsections summarize the different filtering methods applied in this project.

The whole project has been written in Python 2.7, with the code available in: https://github.com/MariolaFG/Master_Thesis.

2.1 Data

The data analyzed in this project has been simulated using haplogen from HaploSim 1.8. package (Software available in: https://git.wageningenur.nl/motaz001/Haplosim).

HaploGenerator can be used separately from HaploSim and combined with Allowrapper to obtain the different haplotypes according to the ploidy level. These tools are described in more detail below.

2.1.1 HaploGenerator

HaploGenerator is a Python pipeline that simulates different haplotypes, given a reference genome and a desired ploidy level. Each haplotype is stored in a FASTA file.

Given a list with different genome locations, HaploGenerator will produce SNPs at those specific positions. It allows using it in random mode, which produces random mutations according to the chosen stochastic model: Poisson or Gaussian.

2.1.2 Allowrapper

The haplotypes created by HaploGenerator are divided according to their subgenome. Using Allowrapper, the haplotypes are combined to produce the same output as if from a single run of HaploGenerator. Allowrapper also produces a text file listing the various types of variants generated by HaploGenerator: homeologous SNPs, homozygous SNPs and heterozygous SNPs.

In the end, we obtain as many FASTA files, containing the haplotypes generated, as the given ploidy level, per genotype.

2.2 Data processing

Once the haplotypes are generated, each FASTA file is processed following the pipeline illustrated in Figure 2.

Reads are generated using ART (ChocolateCherryCake version), which is a tool that simulates NGS reads, mimicking a sequencing technology: Roche's 454, Illumina's Solexa or Applied Biosystems' SOLiD (Huang, W. et al, 2012). In this project,

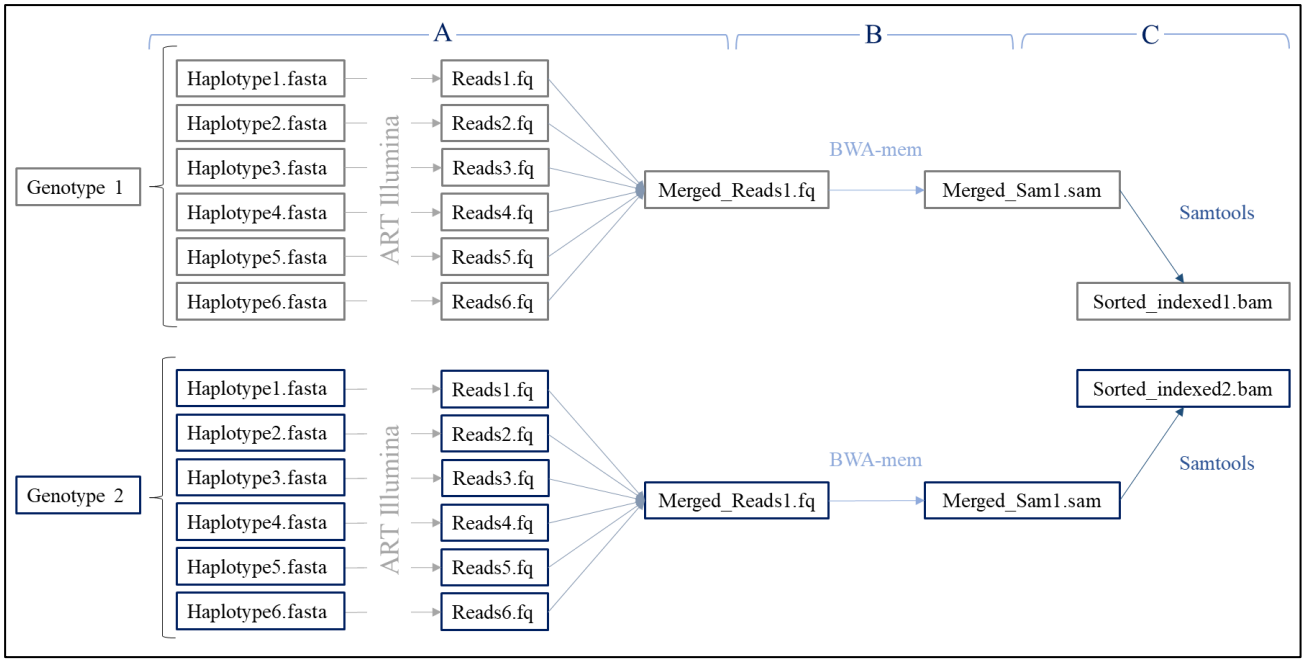


Figure 2. Schematic overview of the data generation pipeline, demonstration on two hexaploidy genotypes.

we used ART Illumina's HiSeq™ 2000 Sequencing System, which primary errors are due to mismatches, to produce single-end reads. Generated read length is 100 bp with a total fold coverage of 120 bp (Figure 2 A).

Reads corresponding to the same genotype are merged into one single FASTQ file.

The merged reads files from all the organisms are mapped to the reference genome given to HaploGenerator using BWA-mem, default parameters, version 0.7.15-r1140 (Li, H. Durbin, R., 2009), yielding Samtools 1.4.1 (Figure 2, B).

Finally, the SAM files from all the genotypes are converted to sorted and indexed BAM files using Samtools 1.4.1 software (Figure 2 C) (Li, H. et al, 2009).

2.3. SWEEP

2.3.1 Original (A)

SWEEP is a Perl script that calls SNPs using Samtools 1.4.1 in combination with Bcftools 1.4.1 and implements a filtering method, a sliding window procedure, to reduce the number of homeologous SNPs. Its operation scheme is illustrated in Figure 3.

First, SWEEP uses "Samtools mpileup", default parameters, to generate BCF files from one or multiple BAM files. It also calculates the normalized Phred scale likelihoods of the possible genotypes, in the following format: RR/RA/AA, where R represents the reference allele and A represents the alternative allele. The most likely genotype is encoded as 0 and least likely genotype as 255.

Once "Samtools mpileup" output is produced, "Bcftools call" performs the actual SNP calling, applying a Bayesian inference to calculate the variant quality (Li, H., 2011), and converts the BCF file to a VCF file.

A sliding window procedure is then applied to the VCF file, consulting each SNP and comparing it with the previous and the next SNPs. SWEEP checks if the genotype

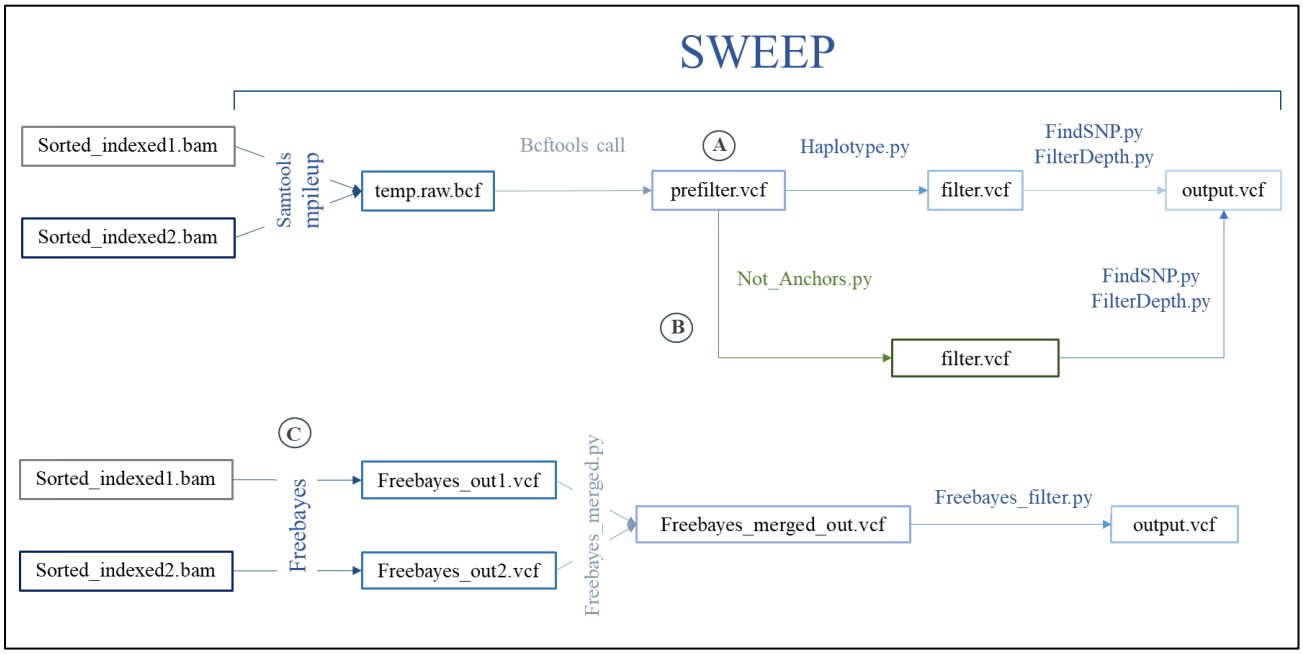


Figure 3. Filtering tools flow chart. A: SWEEP's original pipeline. B: Modified SWEEP's pipeline. C: Freebayes filtering.

likelihoods provided indicate that the SNP under consideration is homozygous to the reference allele in at least one genotype (true allelic SNP) or if, in contrast, it is heterozygous in all genotypes (homeologous SNP).

SWEEP thus selects homologous SNPs by anchoring them with the homeologous SNPs. If it is certain that the previous and the next SNPs are homeologous SNPs at that specific fold coverage and mapping quality of the reads, this provides confidence that the SNP of interest is a homologous SNP, therefore, the false positive rate decreases.

2.3.2 Modified SWEEP (B)

The structure enforced by SWEEP is thus: homeologous SNP – homologous SNP – homeologous SNP. However, some homologous SNPs are not anchored like this, as illustrated in supplementary Figure 1. The more similar the subgenomes are, the lower the number of homeologous SNPs and, therefore, the probability of finding a homologous SNP that is not anchored is higher.

Here, we present a new version of SWEEP, to detect all the homologous SNPS that are not anchored, it does not take the anchors into consideration. It filters out SNPs provided by "Bcftools call" that are heterozygous in both genotypes and select those SNPs that are homozygous to the reference allele in at least one genotype and heterozygous in the rest.

2.4. FreeBayes

FreeBayes is a haplotype-based variant detector, which means that called variants are used to help to detect proximal polymorphisms, therefore, it examines variants in the same context (Garrison, E. Marth, G., 2012). FreeBayes applies a Bayesian statistical method to detect SNPs, indels, multi-nucleotide polymorphisms and complex events. (Garrison, E. Marth, G., 2012).

As illustrated in Figure 3 C, FreeBayes is applied to each of the BAM files corresponding to the different genotypes, applying following parameters: -p "ploidy level" --min-coverage "ploidy level" --min-base-quality 9 --min-mapping-quality 13 --min-alternate-fraction 0,1. Homeologous SNPs are not filtered out by FreeBayes. To select the homologous SNPs, VCF files are merged into one single VCF file, which maintains the probabilities of all the genotypes. This file is then filtered as in the modified SWEEP approach.

Results

3.1 SWEEP's performance decreases at higher ploidy levels and when subgenomes are similar to each other

SWEEP's performance was analyzed for genomes simulated at three different ploidy levels: tetraploid, hexaploid and octoploid, applying three different window sizes: 100 bp (read length), 500 bp and 1000 bp. For all cases, the simulated genome length was 24 kb. SWEEP's performance was also examined and compared in cases where subgenomes are similar to each other, the percentage of homeologous SNPs oscillates from 1% to 3%, and situations where subgenomes are dissimilar, the percentage of homeologous SNPs oscillates from 3% to 5%.

Figure 4 A shows results of SWEEP in tetraploids, this shows that more than 90% of the homologous SNPs were detected, both when subgenomes are similar and when they are dissimilar, Also, the false positive rate is lower than 0,1% and none of the homeologous SNPs were detected as homologous SNPs. A full analysis is presented in supplementary Figure 2.

Figure 4 B illustrates that SWEEP's performance in hexaploids in dissimilar subgenomes is better than in the case where the subgenomes are similar. Even though this difference in performance is not that significant, it is higher than the difference in tetraploids. False positive rate slightly increases compared to previous cases. A full analysis is presented in supplementary Figure 3.

As shown in Figure 4 C, SWEEP's homologous SNP detection in octoploids decreases significantly considering the previous cases, where more than 90% of homologous SNPs were detected. In dissimilar subgenomes the homologous SNP detection rate drops to 79% and in similar subgenomes, hardly 70% of the homologous SNPs are found.

Also, the false positive rate increases, being almost 3% in dissimilar subgenomes, see full analysis in supplementary Figure 4.

The higher the ploidy level, the more subgenomes are present in the genotype, therefore, calculation of the genotype likelihoods for homeologous SNPs is harder. In tetraploids, if there is a homeologous SNP in a specific position, half of the reads will contain the alternate allele. In octoploids, however, 25% of the reads will contain the alternate allele. This is the reason why in

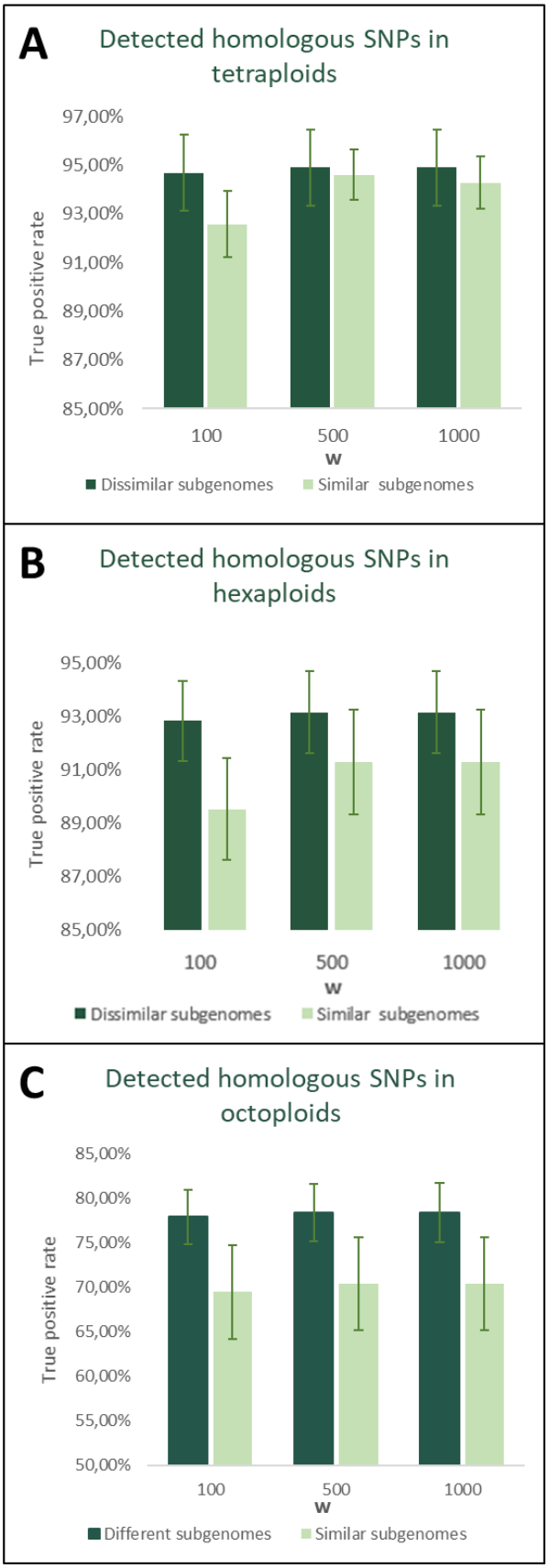


Figure 4. Homologous SNP rate detected by SWEEP in dissimilar subgenomes (dark green) and in similar subgenomes (light green), using three window sizes (w). A: tetraploids, B: hexaploids, C: octoploids. Error bars indicate standard deviation of ten experiments.

octoploids SWEEP does not manage to filter out all the homeologous SNPs.

3.2 Modified SWEEP outperforms SWEEP

Modified SWEEP and SWEEP's performance was analyzed and compared following the same methodology as described in the previous section, only for hexaploids and octoploids.

Figure 5 (A, B) shows results of the different filtering tools in hexaploids, this shows that at this level of ploidy, SWEEP's original code detects 93% of the true allelic SNPs in the case with the subgenomes are dissimilar and 91% in the case where the subgenomes are similar. In modified SWEEP, in which anchors are not included, performance increases, while the false positive rate stays the same as when using SWEEP's original code. Full analysis is included in supplementary Figure 5.

Results for the performance of modified SWEEP and SWEEP in octoploids are illustrated in Figure 5 (C, D). As seen in the former section, at higher ploidy levels, SWEEP does not detect as many homologous SNPs as in lower ploidy levels. Using modified SWEEP, slightly improves homologous SNP detection without raising the false positive rate, compared with SWEEP, however performance hardly reaches 80% when subgenomes are dissimilar and 73% when subgenomes are similar.

Modified SWEEP considers a homologous SNP every SNP that is homozygous to the reference allele in at least one genotype and

heterozygous in the rest. If Bcftools provides all homologous SNPs and genotypes likelihoods are correctly calculated, the new SWEEP version should be able to detect all the homologous SNPs, which indicates that Bcftools does not provide all homologous SNPs.

3.3 Filtered FreeBayes detects more homologous SNPs in high ploidy levels than SWEEP and modified SWEEP

As illustrated in Figure 5 (A, B), in hexaploids, filtered FreeBayes detects more than 90% of homologous SNPs, regardless subgenome similarity. However, it does not detect as many homologous SNPs as modified SWEEP. False positive rate is higher and the percentage of homeologous SNPs that were not filter out is also higher compared with SWEEP and modified SWEEP.

Filtered FreeBayes is used separately for each genotype and then merged, so filtered FreeBayes error rate will be obtained per each genotype, which explains false positive increase compared with SWEEP and modified SWEEP. This could also cause that not all homeologous SNPs are filtered out, because for this method it is necessary that filtered

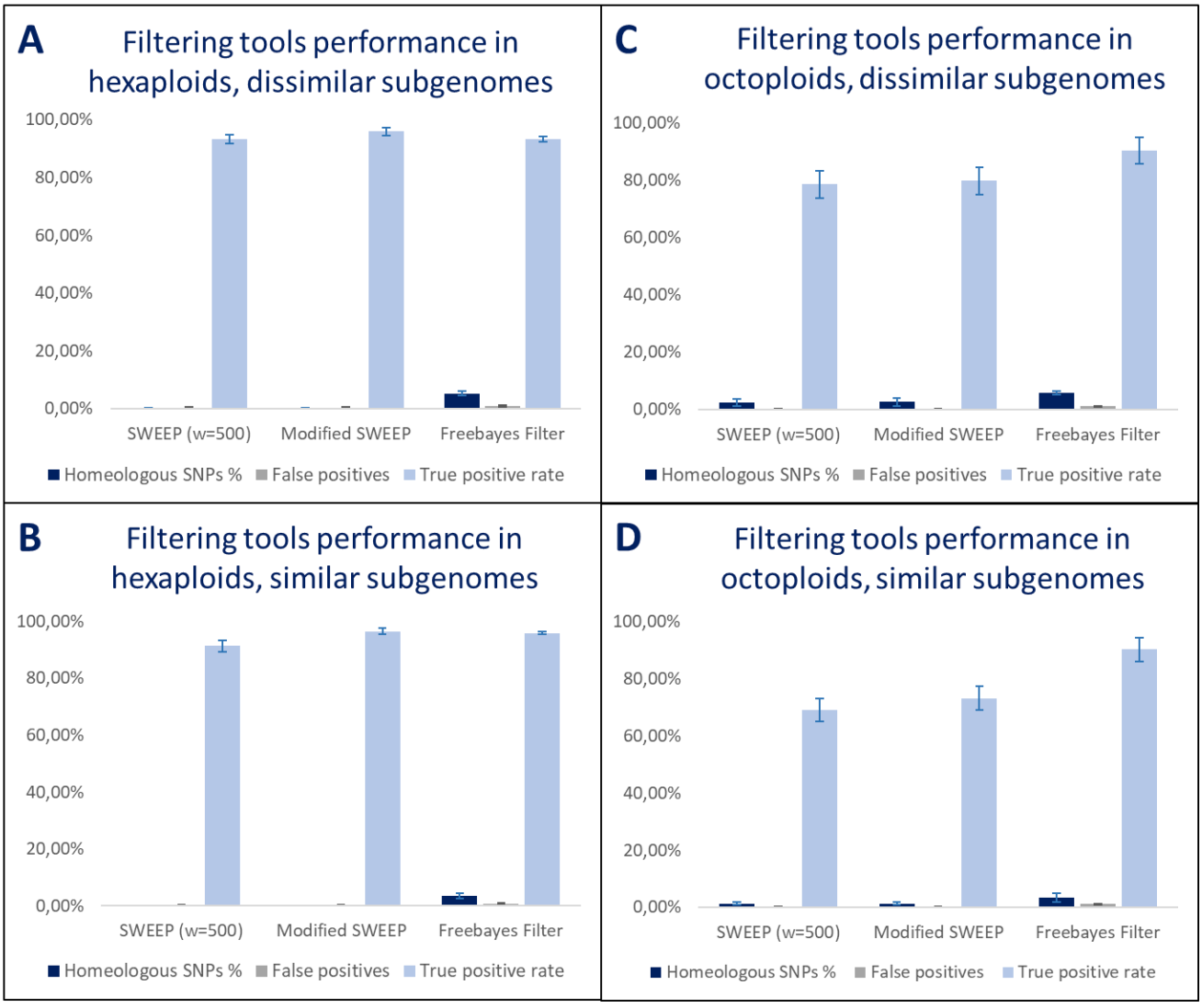


Figure 5. Comparison of the filtering tools performance in hexaploids (A, B) and octoploids (C, D), in dissimilar subgenomes (A, C) and in similar subgenomes (B, D). Error bars indicate standard deviation of ten experiments.

FreeBayes detects homeologous SNPs in all the genotypes, if for a specific position there is a homeologous SNP and filtered FreeBayes only detects it in one genotype, it will be called as a homologous SNP.

In contrast, in octoploids, filtered FreeBayes detects more than the 90% of homologous SNPs, regardless subgenome similarity, as illustrated in Figure 5 (C, D), so it outperforms SWEEP and modified SWEEP in homologous SNP detection in higher ploidy levels. False positive rate is higher than the rest of the filtering tools, and, as explained before, not all homeologous SNPs were filtered out, however, SWEEP's original code does not filter out all homeologous SNPs either, see supplementary Figures 5 and 6.

3.4 Bcftools does not detect all homologous SNPs in high ploidy levels

As illustrated in Figure 6, Bcftools only returns approximate 80% of homologous SNP when subgenomes are dissimilar and 73% when the subgenomes are similar. In other words, SWEEP performance is fully limited by prior SNP calling.

This might be due to the presence of heterozygous SNPs, which Bcftools does not detect as well as homozygous SNPs when the ploidy level is high, as can be seen in Figure 6.

Bcftools may not detect heterozygous SNPs because it is operating on a single position at a time, which makes it highly dependent on the reads. When analyzing two octoploid genotypes, if there is a heterozygous SNP in a specific position, it will be present just in 12,5% of the reads. Freebayes is a haplotype-based variant detector, it applies information from nearby possible polymorphisms to estimate the haplotype, which helps to detect rare variations, such us heterozygous SNPs, regardless the ploidy level.

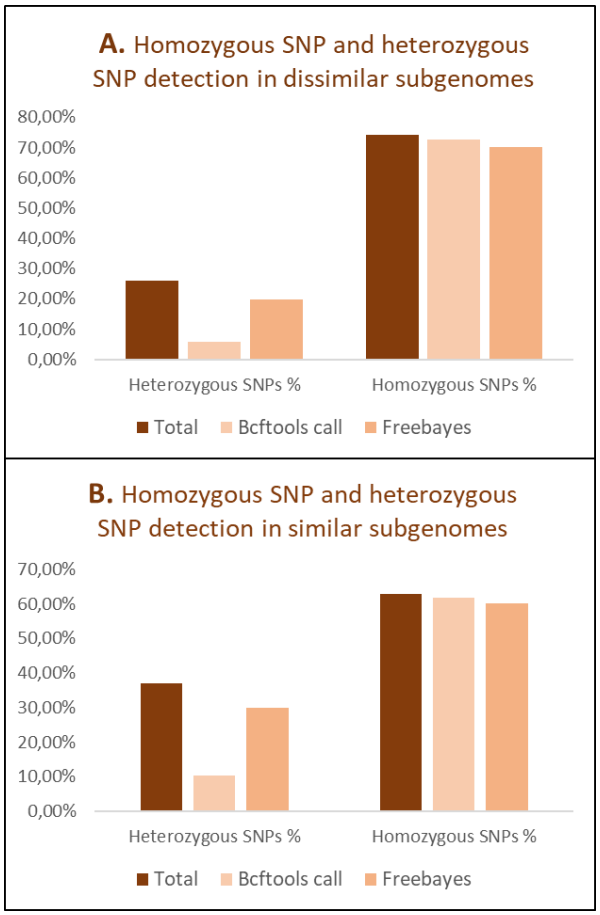


Figure 6. Homozygous SNP and heterozygous SNP detection in octoploids by Bcftools and Freebayes, compared to the total amount of homologous SNPs in the genome.

3.5 The performance Bcftools decreases with coverage

To test if a total coverage of 120 was not optimal for Bcftools's perfomance, a fold coverage screening was performed.

As illustrated in Figure 7, Bcftools is highly dependent on coverage, and therefore, so is SWEEP. There is a decrease in performance when fold coverage increases, this is due to

Samtools. When there are too many reads, Samtools picks 255 reads randomly to calculate genotype likelihoods (Li, H., 2010), sampling heterozygous SNPs in the reads is harder because they are in less proportion on the reads, so homologous SNP detection decreases.

FreeBayes, in contrast, slightly increases with coverage, however, it is always between 85% to 95%, regardless subgenome similarity.

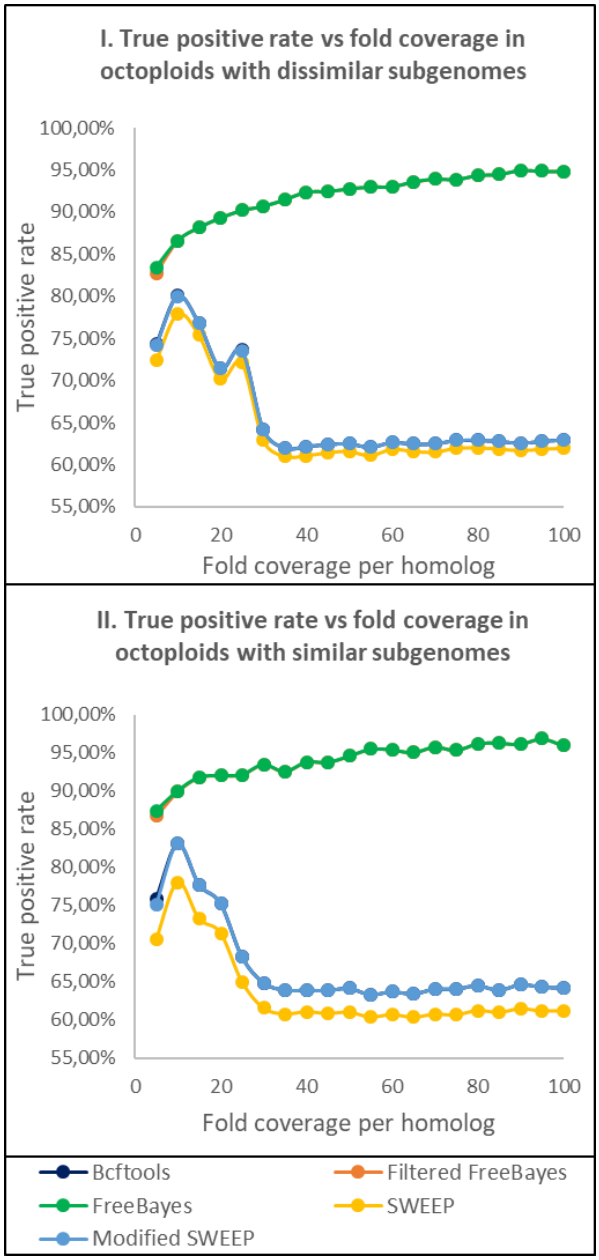


Figure 7. True positive rate obtained by different SNP calling tools, before and after filtering.

Previous experiments were run in a total coverage of 120 (coverage per homolog of 15), which is not the optimal coverage for SWEEP in octoploids, as illustrated in Figure 7, however, at the optimal coverage, Bcftools hardly reaches 80%-85%, depending subgenome similarity.

Discussion

In this project we tested the performance of SWEEP, a tool to detect homologous SNPs in allopolyploids. A pipeline was built to simulate NGS data from allopolyploid organisms with three different ploidy levels (tetraploid, hexaploid and octoploid). Reads were generated from this simulated NGS data and then mapped to the reference genome.

SWEEP's accuracy relies on the genotype likelihoods given by Samtools and Bcftools, which may not be calculated correctly at higher levels of ploidy due to the presence of more subgenomes.

Moreover, SWEEP uses homeologous SNPs as anchors to detect homologous SNPs. However, the more similar subgenomes are to each other, the fewer homeologous SNPs will be present, and, therefore, the more likely it is to find homologous SNPs that are not anchored. According to (Clevenger, J. et al, 2015), subgenomes can exhibit more than 3% divergence, but there are organisms, such as peanut, where subgenomes are too similar. This simulation includes analysis in similar subgenomes (1% to 3% divergence) and dissimilar subgenomes (3% to 5% divergence).

SWEEP's performance in tetraploids is almost perfect. More than 90% of homologous SNPs are detected with a false positive rate lower than 0,1%. Homeologous SNPs are filtered out, regardless the similarity within the subgenomes. Therefore, SWEEP is a good filtering tool for tetraploids. At ploidy levels 6 and 8, however, SWEEP performance was not perfect; the false positive rate increased, not all homeologous SNPs were filtered out and the percentage of homologous SNPs detected decreased. In octoploids, the homologous SNP detection rate hardly reached 80% when subgenomes were dissimilar to each other and 70% in the situation where subgenomes are similar to each other.

We then investigated modifications to SWEEP to address these problems.

The first is not using the anchors at all, by just selecting SNPs that are homozygous to the reference allele in at least one genotype and heterozygous in other genotypes. In this way, all the homologous SNPs that SWEEP would filter out just because they are not anchored, will be detected. Therefore, we expected that the homologous SNP detection rate would be higher than in SWEEP's original code, especially when subgenomes are similar to each other. Although the modification resulted in a slight improvement, the true positive rate was lower than expected, which suggested that Bcftools was not returning all homologous SNPs.

We showed that at high ploidy levels, this is indeed the case. At higher ploidies and at high coverage, Bcftools especially fails to detect heterozygous SNPs. This may be because Bcftools acts at single site level, which makes it highly dependent on the reads. Heterozygous SNPs are present in a small proportion in the reads in polyploids, making it harder to detect them.

It is important to note that the modification did not significantly increase the false positive rate, which implies that SWEEP's use of homeologous SNPs as anchors may not be needed.

A second approach was to use FreeBayes, a haplotype-based variant detector. Because of its use of information from proximal reference-relative variations, we expected it not to be as dependent on the reads as Bcftools and to be able to detect more homologous SNPs at high ploidy levels. FreeBayes detects both homeologous SNPs and homologous SNPs, so we developed a Python pipeline to filter out the homeologous SNPs, selecting all SNPs that are not present in all the genotypes. This approach managed to detect more than the 90% of homologous SNPs in hexaploids and octoploids. However, the false positive rate was higher than the one obtained from SWEEP, and not all homeologous SNP were filtered out. The higher false positive rate is caused by FreeBayes being used for each genotype, which, multiplies by the number of genotypes FreeBayes original error rate. If FreeBayes fail to detect a homeologous SNP in one genotype, it is detected as a homologous SNP in further filtering, which explains why not all homeologous SNPs were filtered out.

Conclusions & Future work

The use of homeologous SNPs as anchors SNPs reduces true positive rate and may not be needed. SWEEP detects homeolgous SNPs and uses them to relate homologous SNPs into the same context, so it reduces false positive rate (Clevenger, J. et al, 2015). However, the SWEEP modification proposed here, which does not use homeologous SNPs as anchors, does not have a significant false positive rate and the true positive rate slightly improves. This research could be further validated by using real data as using real data, as the simulation used in this project only contained mismatches, but, insertions and deletions also occur, and it may alter the false positive rates obtained here.

SWEEP performs well at low ploidy levels, but at high ploidy levels it is limited by problems in prior SNP calling. According to (Clevenger, J. et al, 2015), SWEEP false positive rate ranges from 0,01% to 0,02% and it returns almost all homologous SNPs. This study was performed in tetraploids and was also proved in this project. However, we showed here that at high ploidy level, SWEEP does not perform as well as in tetraploids, so the future research in a real data discussed before, should be of organisms of high ploidy level.

Samtools and Bcftools combination calculate single site frequencies, relying on the reads to do so (Li, H., 2011), this causes a drop on the true positive rate at high ploidy levels, so would be useful to investigate new SNP calling tools for prior SNP calling in SWEEP.

FreeBayes is a good approach for prior SNP calling before applying SWEEP filtering at high ploidy levels. FreeBayes improves genotyping accuracy for rare variations using information from nearby polymorphisms (Garrison, E., 2012), which allows it to return more than 90% of homologous SNPs, regardless the ploidy level or subgenome similarity. The filtering step proposed here after SNP calling by FreeBayes does not filter out all homeologous SNP, so further research on how to improve the filtering step would be useful. SWEEP's filtering method, manages to filter almost all homeologous SNPs (Clevenger, J., 2015), so combining FreeBayes with SWEEP could be a good method to detect homologous SNPs in allopolyploids at high ploidy levels.

References

Altshuler, D. Brooks, L.D. Chakravarti, A. Collins, F.S. Daly, M.J. Donnelly, P. (2005). *A haplotype map of the human genome*. International HapMap Consortium. Nature 437, pp. 1299-1320.

Clevenger, J. Chavarro, C. Pearl, S.A. Ozias-Akins, P. Jackson, S.A. (2015). *SNP identification in polyploids: a review, example and recommendations*. Mol. Plant.

Clevenger, J. Ozias-Akins, P. *SWEEP: A tool for filtering high-quality SNPs in polyploid crops*. (2015). G3 (Bethesda) 5, 1797–1803.

Encyclopaedia Britannica. (2017). *Genetic marker*. Encyclopaedia Britannica 15th ed., s.v.

Garrison, E. Marth, G. (2012). *Haplotype-based variant detection from short-read*

sequencing. arXiv Prepr arXiv12073907 [q-bio.GN].

Griffiths, A.J.F. Miller, J.H. Suzuki, D.T. Lewontin R.C. Gelbart, W.M. (2000). *An Introduction to genetic analysis*. San Francisco: W.H. Freeman. ISBN-10: 0-7167-3520-2. p 484-488.

Hirakawa, H. Shirasawa, K. Kosugi, S. Tashiro, K. Nakayama, S. Yamada, M. Kohara, M. Watanabe, A. Kishida, Y. Fujishiro, T. Tsuruoka, H. Minami, C. Sasamoto, S. Kato, M. Nanri, K. Komaki, A. Yanagi, T. Guoxin, Q. Maeda, F. Ishikawa, M. Kuhara, S. Sato, S. Tabata, S. Isobe, S.N. (2014). *Dissection of the Octoploid strawberry genome by deep sequencing of the genomes of Fragaria species*. DNA Res. 21, 169–181.

Huang, W. Li, L. Myers, J. Marth, G. (2012). *ART: a next-generation sequencing read simulator*. Bioinformatics 28 (4): 593-594.

Jiang, G.L. (2013). *Molecular Markers and Marker-Assisted Breeding in Plants*. Plant Breeding from Laboratories to Fields ed. Andersen S. B. DOI: 10.5772/52583 InTech.

Kwok, P.Y. Chen, X. (2003). *Detection of single nucleotide polymorphisms*. In Genetic Engineering, Principles and Methods, ed. JK Setlow. New York: Plenum. 20:125–34.

Li, H. (2010). *Mathematical Notes on SAMtools Algorithms*. http://lh3lh3.users.sourceforge.net/download/samtools.pdf.

Li, H. (2011). *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data*. Bioinformatics, 27(21): 2987–2993.

Li, H. Durbin, R. (2009) *Fast and accurate short read alignment with Burrows-Wheeler Transform*. Bioinformatics, 25:1754-60. [PMID: 19451168].

Li, H. Handsaker, B. Wysoker, A. Fennell, T. Ruan, J. Homer, N. Marth, G. Abecasis, G. Durbin, R. (2009). *The Sequence Alignment/Map format and SAMtools*. Bioinformatics. 25, S. 2078.

Shaw G. (2013). *Polymorphism and single nucleotide polymorphisms (SNPs)* [J]. BJU Int, 112(5): 664-5.

Supplement

Figure 1

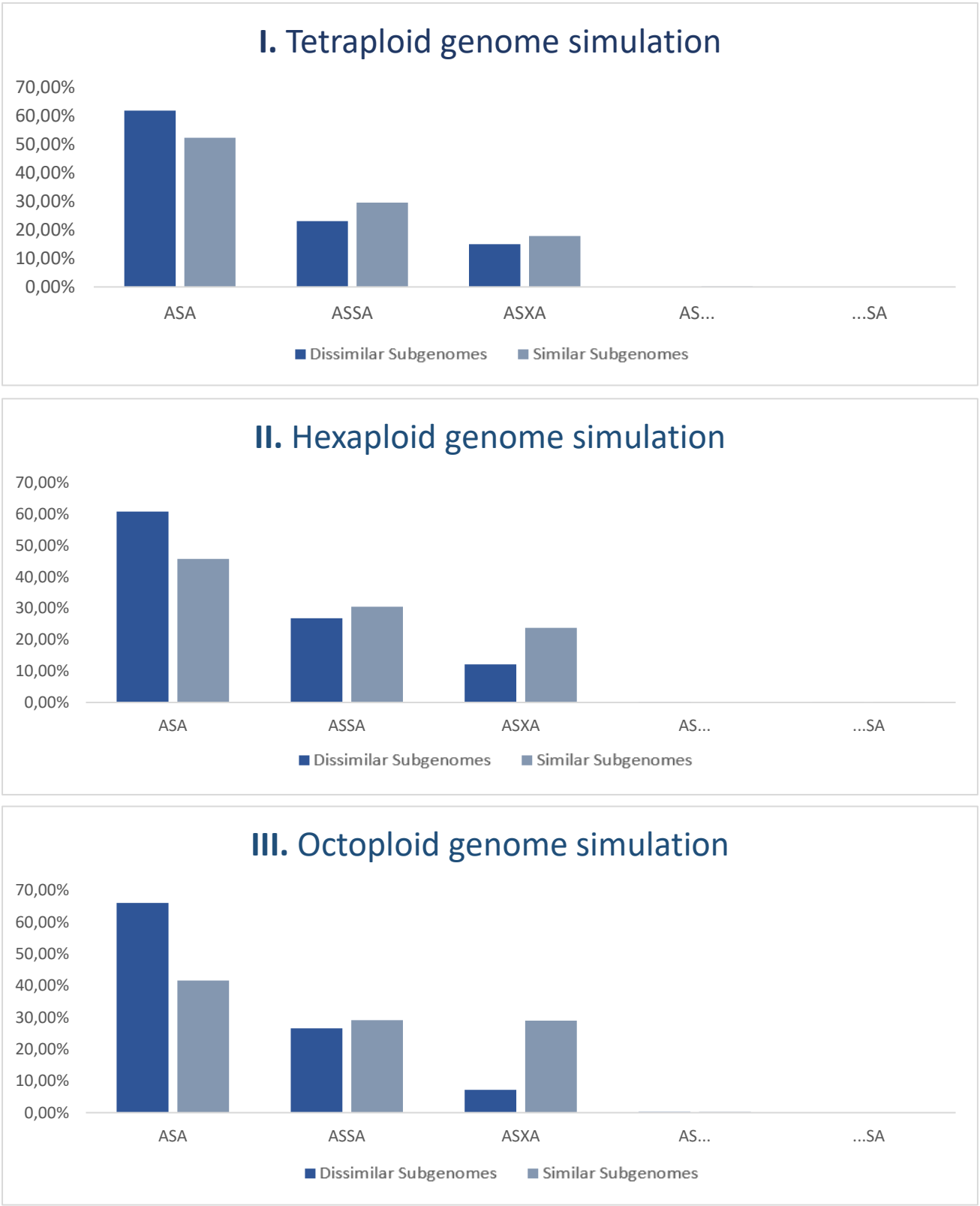
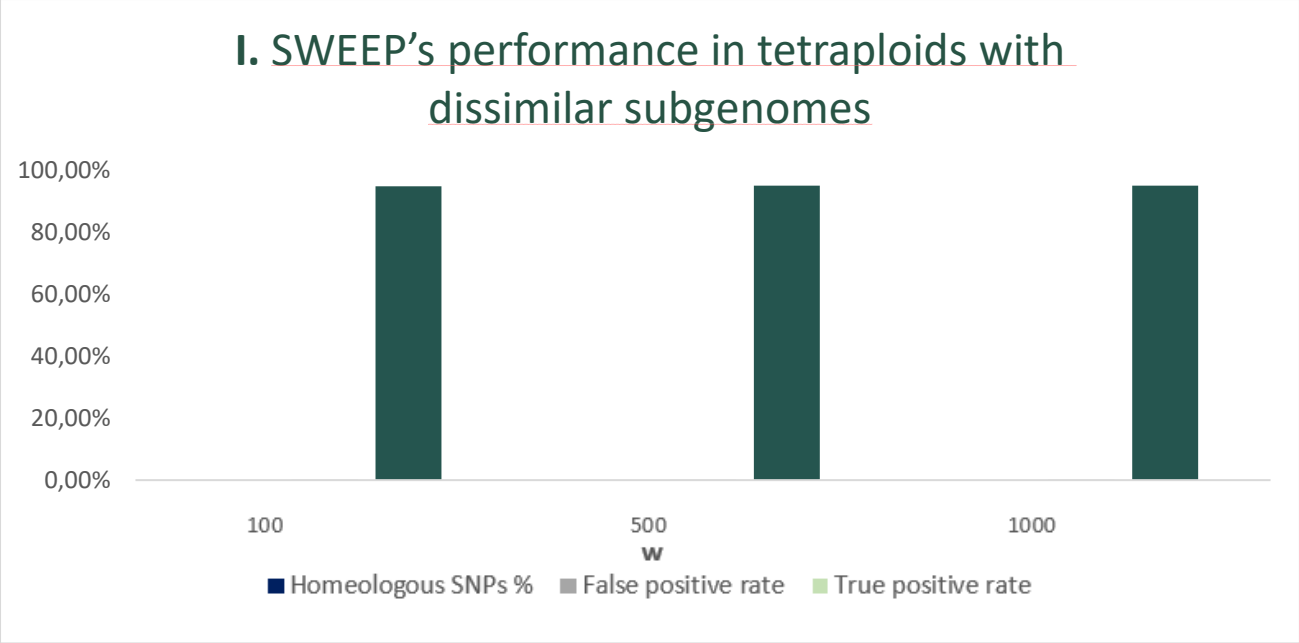


Figure 1. SNP location frequency in tetraploids (I), hexaploids (II) and octoploids (III) in simulation. A = Anchor SNP (homeologous SNP), S = Homologous SNP, X = More than two true allelic SNPs.

Figure 2



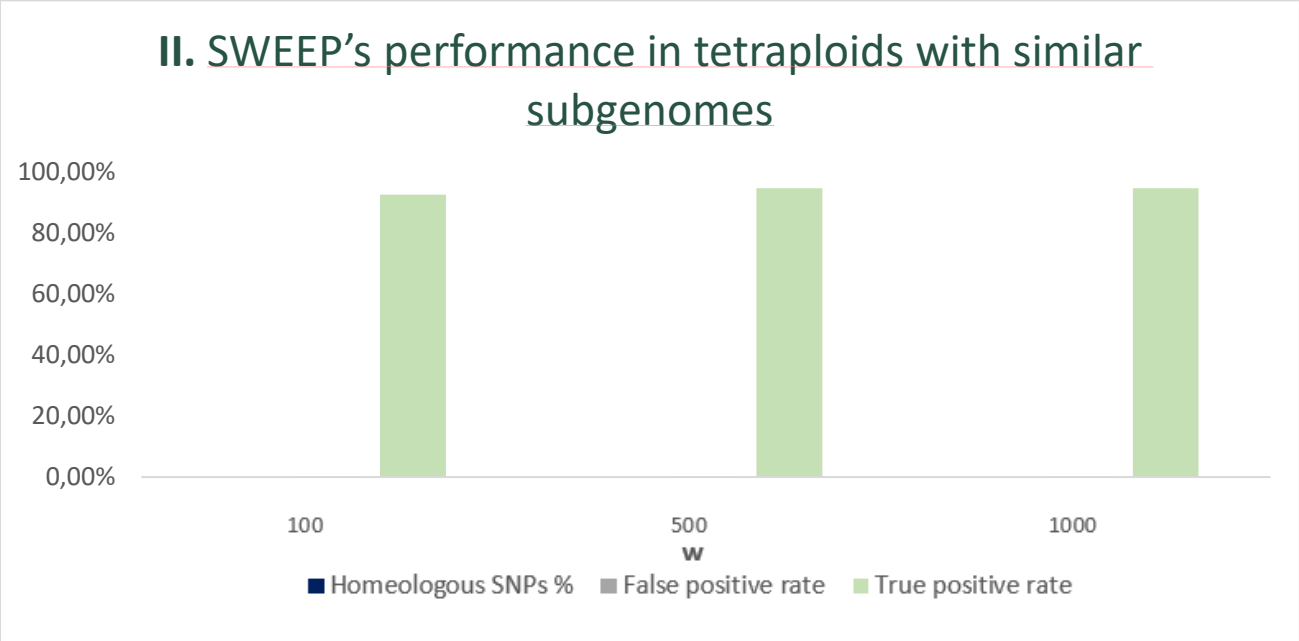


Figure 2. SWEEP's performance in tetraploids, using three window sizes.

Figure 3

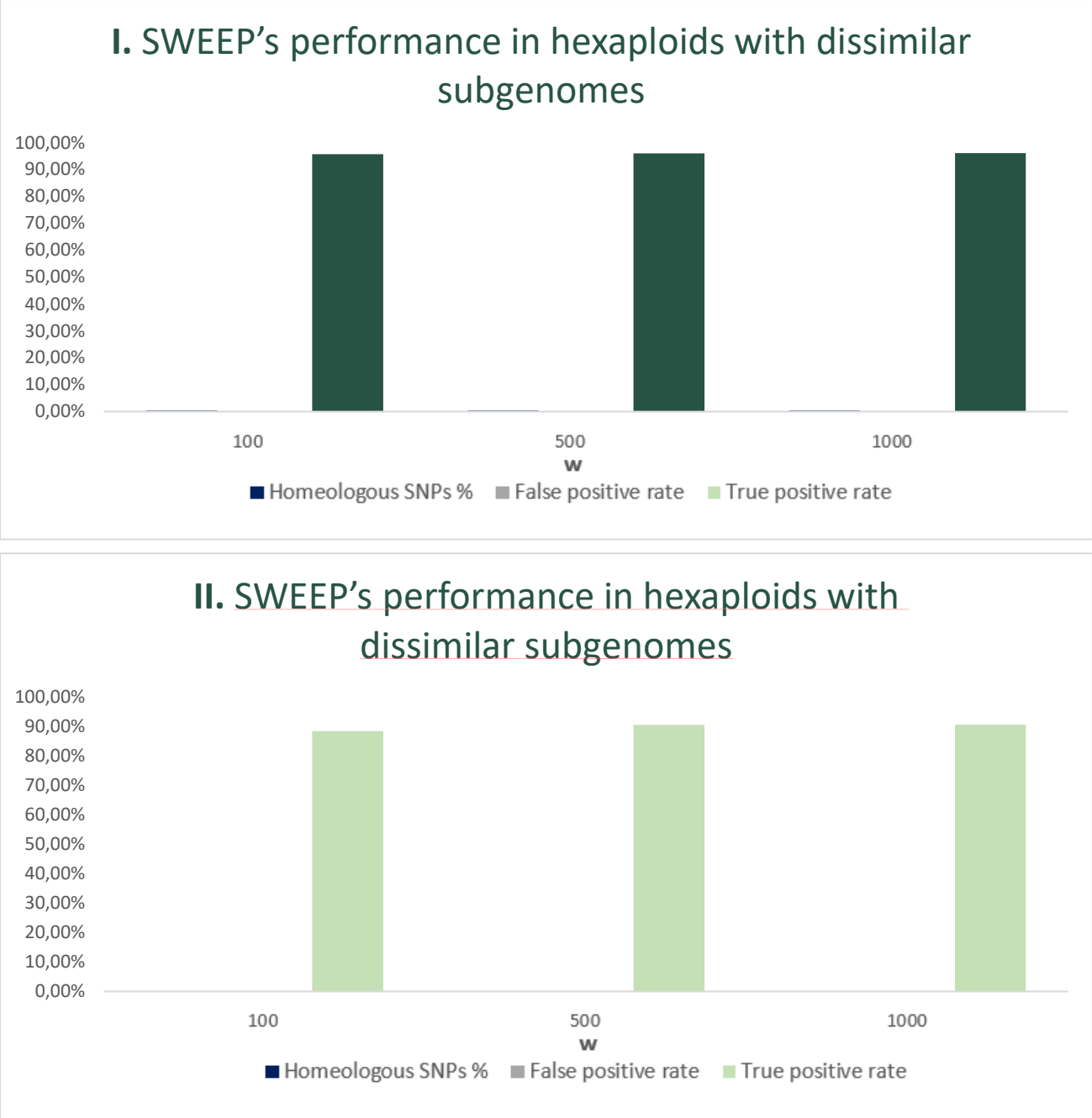


Figure 3. SWEEP's performance in hexaploid, using three window sizes.

Figure 4

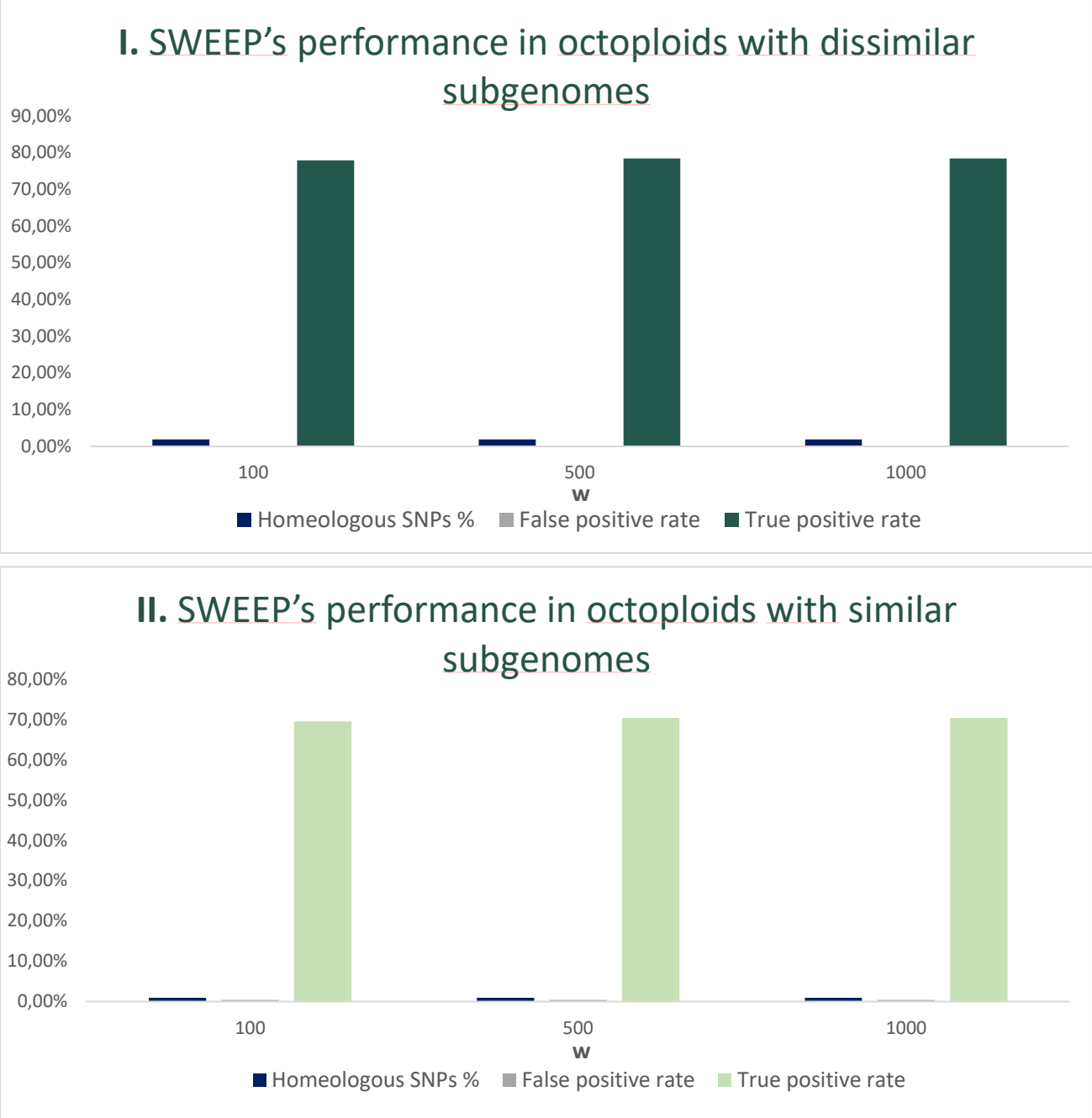


Figure 4. SWEEP's performance in octoploid, using three window sizes.

Figure 5

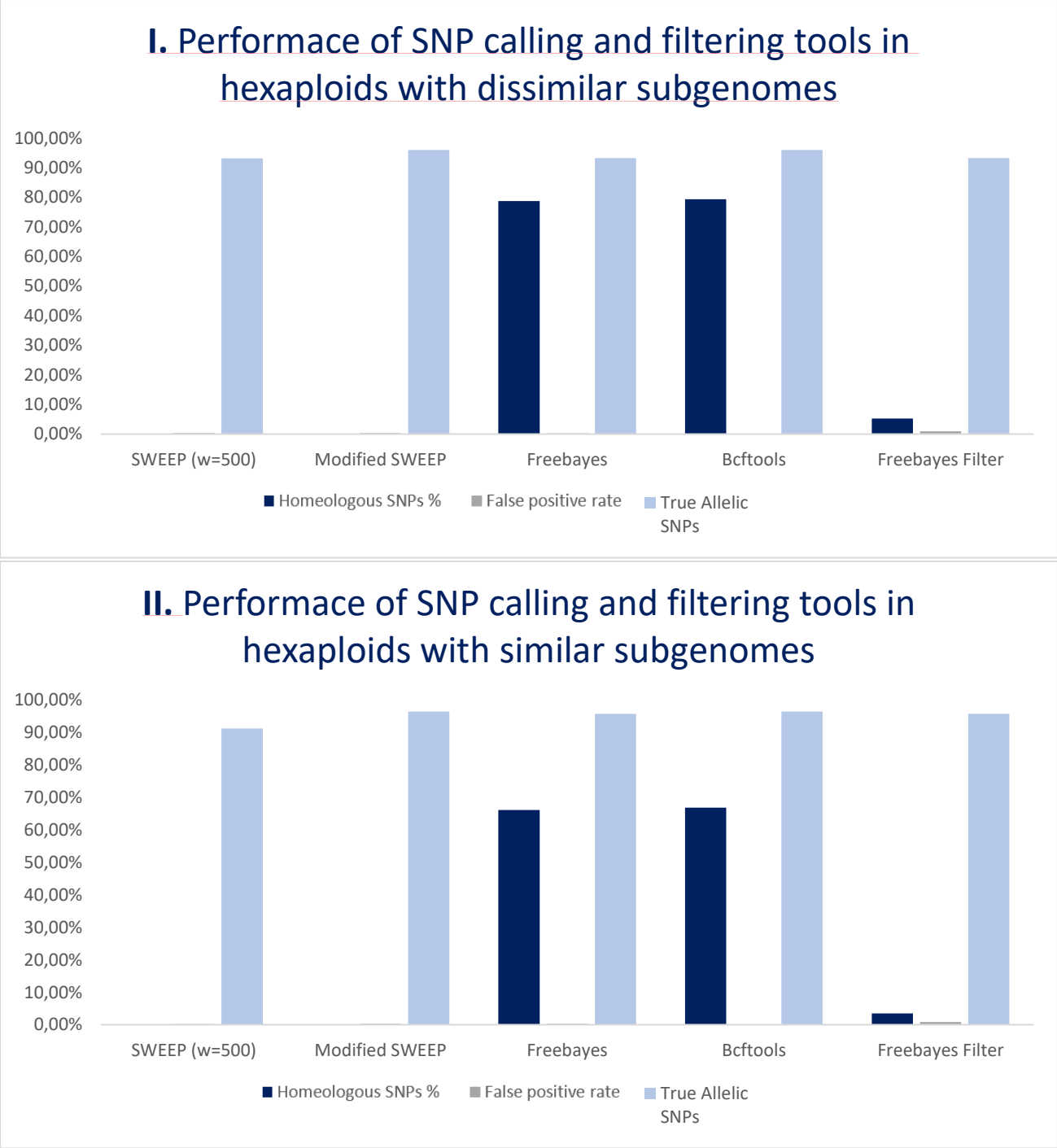


Figure 5. Filtering tools performance in SNP detection in hexaploids with dissimilar subgenomes (I) and similar subgenomes (II).

Figure 6

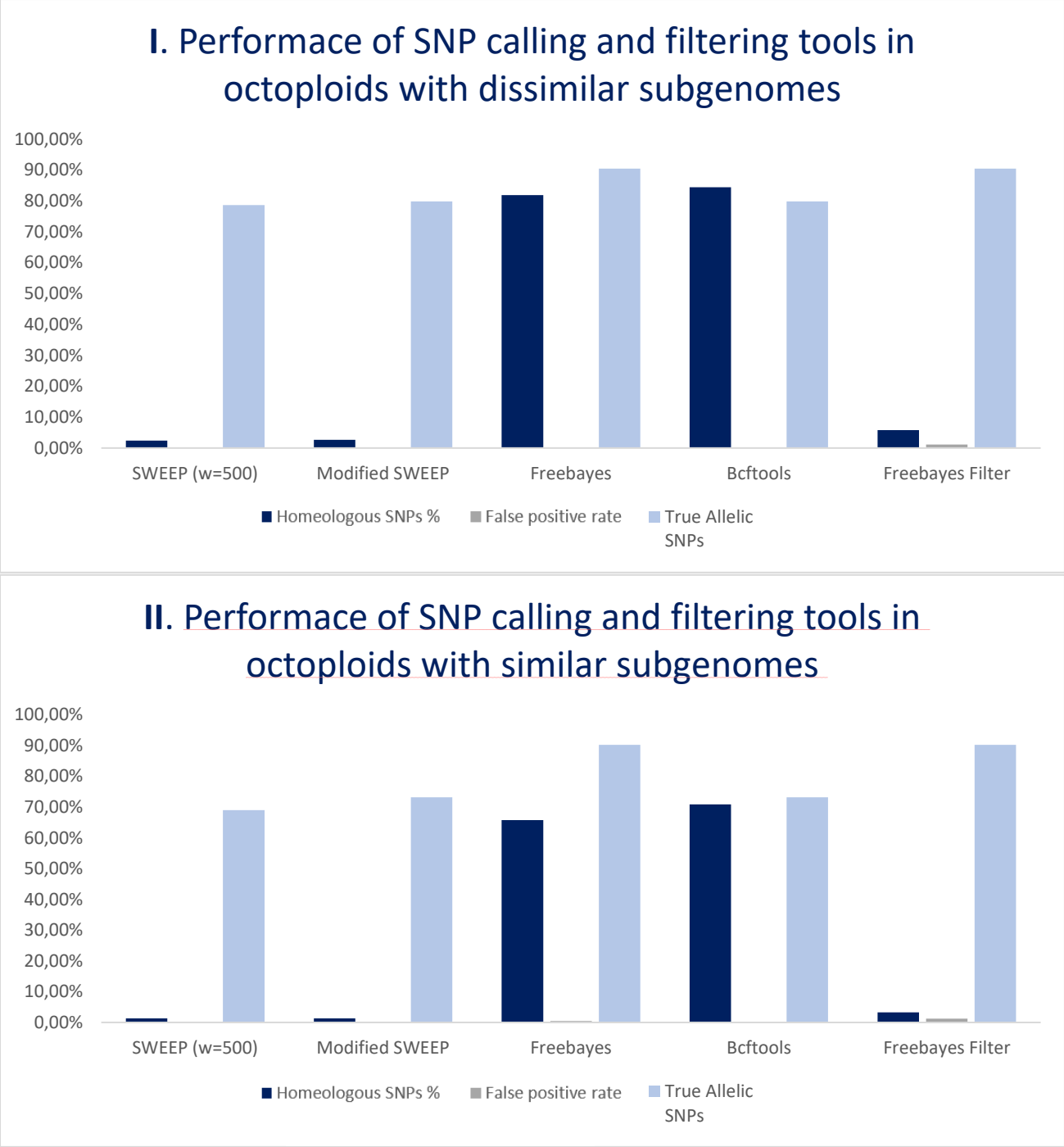


Figure 6. Filtering tools performance in SNP detection in octoploids with dissimilar subgenomes (I) and similar subgenomes (II).