Abstract.md

Identification of genomic regions carrying a causal mutation in unordered genomes.

Whole genome sequencing using high-throughput sequencing (HTS) technologies offers powerful opportunities to study genetic variations. Mapping the mutations responsible for phenotypes is generally an involved and time-consuming process so researchers have developed user-friendly tools for mapping-by-sequencing, yet they are not applicable to organisms with non-sequenced genomes.

We introduce SDM (SNP Distribution Method), a reference independent method for rapid discovery of mutagen-induced mutations in typical forward genetics screens. SDM aims to order a disordered collection of HTS reads or contigs so that the fragment carrying the causative mutations can be identified. SDM uses typical distributions of homozygous SNPs that are linked to a phenotype-altering SNP in a non-recombinant region as a model to order the fragments. To implement and test SDM, we created model genomes with SNP density based on *Arabidopsis thaliana* chromosome and analysed fragments with size distribution similar to reads or contigs assembled from HTS sequencing experiments. SDM groups the contigs by their normalised SNP density and arranges them to maximise the fit to the expected SNP distribution. We analysed the procedure in existing data sets by examining SNP distribution in recent out-cross [@Galvão et al. 2012], [@Uchida et al. 2014] and back-cross experiments [@Allen et al. 2013], [@Monaghan et al. 2014] in *Arabidopsis thaliana* backgrounds. In all the examples we analysed, homozygous SNPs were normally distributed around the causal mutation. We used the real SNP densities obtained from these experiments to prove the efficiency and accuracy of SDM. The algorithm succeed in the identification of the genomic regions of small size (10-100 kb) containing the causative mutations.

Background.md

Background

Forward genetic screens have been a fundamental strategy to find genes involved in biological pathways in model species. In these a population is treated with a mutagen that alters the DNA of individuals in some way, e.g. induction of guanine-to-adenine substitutions using ethylmethane sulfonate (EMS) [@Page:2002], then individuals with a phenotype of interest are isolated from a mutagenized population and a recombinant mapping population is created by back-crossing to the parental line or out-crossing to a polymorphic ecotype [@Etherington:2014]. The recombinant population obtained from that cross will segregate for the mutant phenotype and individuals showing the mutant phenotype will carry the causal mutation, even if the genomic location is unknown. The recombination frequency between the causal mutation and nearby genetic markers is low, so the alleles of these linked genetic markers will co-segregate with the phenotype-altering mutation while the remaining unlinked makers segregate randomly in the genome [@Schneeberger:2014aa]. Hence, the allele distribution analysis can unhide these low recombinant regions to identify the location of the causal mutation. This process of genetic analysis is often referred to as bulk segregant analysis (BSA) [@Michelmore:1991aa].

Traditional genetic mapping is a work intensive and time-consuming process but recent advances in high-throughput sequencing (HTS) have greatly accelerated the identification of mutations underlying mutant phenotypes in forward genetic screens. Several methods as SHOREmap [@Schneeberger:2009], [@Sun:2015], NGM [@Austin:2011] or CandiSNP [@Etherington:2014] based on bulked-segregant analysis of F2 progeny have succeed in the mutant identification in *Arabidopsis thaliana*. All these methods depend on an assembled reference genome and cannot be used in species for which a reference genome is not available. Some alternative solutions as using reference sequences of related species have been proposed [@Wurtzel:2010], [@Livaja:2013aa], but these need low sequence divergence and high levels of synteny between the mutant reads and the related reference sequence and this has restained the application of this approach [@Schneeberger:2014aa],[@Nordstrom:2013aa].

Substantial effort is being made to sequence many species but reasonable completion of a sequence remains expensive and time-comsuming, and fragmented draft genomes present certain limitations in use for mutation mapping in many circumstances. Fast-evolving and repetitive genes such as disease resistance genes [@Song:2003aa] might be absent or divergent from draft reference genome assemblies. Also, draft genomes often have gaps that can frustrate alignments.

In the last few years, several reference-free methods for general mutation identification have been proposed [Iqbal:2012aa], [Nordstrom:2013aa], [@Minevich:2012aa], [@Abe:2012] to solve the reference sequence restriction, but none have been extended to allow for direct identification of causative mutations. [@Abe:2012], [@Takagi:2015aa], [@Schneeberger:2014aa].

We propose SDM, a fast causative mutant identification method based on a simple reference-free contig assembly that allows the detection of candidate causative SNPs. Instead of relying on a genome comparison, we focus on the SNP linkage around the causal mutation and analyse the SNP distribution to identify the chromosome area where the putative mutated gene is located. SDM does not rely on previously known genetic markers and can be used on extremely fragmentary genome assemblies, even down to the level of long reads.

Methods.md 2. Methods ===

2.1. Model genome generation

We used model genomes to develop our mutant identification method. These were created by asigning an idealised SNP distribution to a set of randomly shuffled sequences that imitate contigs. We created different model genomes based on the 34.9 Mb of *Arabidopsis thaliana* chromosome I. The FASTA sequence used was TAIR10_chr1.fas from The Arabidopsis Information Resource [@Lamesch:2012aa]. *Arabidopsis thaliana* makes an ideal model genome because it is small, the genetic variation is well-described and it contains a small content of repeats.

To create the model genomes, we used different variations of the script https://github.com/edwardchalstrey1/fragmented_genome_with_snps/blob/master/create_model_genome.rb. A detailed protocol and the code to replicate the model genomes were deposited in the GitHub repository https://github.com/pilarcormo/SNP_distribution_method/tree/master/Small_genomes.

In the model genome, homozygous SNPs follow a normal distribution (as proven in the section 3.4). The R function rnorm (n, mean, sd) was used to define the homozygous SNP distribution. The mean was specified before running the script in the middle of the model genome, generating a normal distribution with equally sized tails. The standard deviation (sd) was 2 times the mean value. Heterzygous SNPs followed a uniform distribution across the genome length. The R function runif (n, min, max) was used to define the heterozygous SNPs. The min value was fixed to one and the max value was the model genome length. For both functions, n varied in each genome to meet the requirement of finding a SNP every 500 bp.

A minimum contig size is provided as an argument when running the script, and the maximum contig size is obtained doubling the minimum value. Contig size randomly oscillates between these 2 values.

First, we ran small_model_genome.rb to create 1, 3, 5, 7, 11, 13 and 15 Mb genomes with 1 SNP every 500 bp and 2 different contig sizes (1300 and 700 bp). Each genome was replicated 5 times, making a total of 70 genomes which can be found at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Small_genomes/arabidopsis_datasets/1-15Mb.

Then, we ran chr1_model_genome.rb to use the whole chromosome I length to generate longer model genomes. A more realistic SNP density was used for these models (1 SNP every 3000 bp). In this case, 3 contig sizes were employed (1000, 2000 and 4000 approximately) and we replicated each model 5 times, obtaining 15 model genomes more. Those were deposited at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Small_genomes/arabidopsis_datasets/30Mb under the names chr1_i for 1000 contigs, chr1_A_i for 2000 contigs and chr1_B_i for 4000 contigs genomes.

2 sets of model genomes with a non-centered mean were also generated to test SDM filtering step. These genomes were divided in 2000 contigs. They can be found at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Small_genomes/arabidopsis_datasets/30Mb under the names chr1_C_i, which presents an approximated 20% deviation to the right, and chr1_E_i, which presents an approximated 20% deviation to the left.

Model genomes folders contain a FASTA file with the correct fragment order, a FASTA file with the randomly shuffled fragments and a VCF file with the homozygous and heterozygous SNP positions. For simplicity, homozygoys SNPs are given a fixed Allele Frequency (AF) of 1 and heterozygous SNPs are given an AF of 0.5.

2.2. SDM implementation using model genomes

Fig.1 shows SDM workflow.

The first step in the SDM pipeline is the homozygous to heterozygous SNPs ratio calculation on each contig. The ratio of homozygous to heterozygous SNPs on a contig n is defined as the sum homozygous SNPs on n plus 1 divided by the sum of heterozygous SNPs on n plus 1:

$$Ratio_n = \frac{(\sum Hom)+1}{(\sum Het)+1}$$

Then, the effect of contig length on SNP density is reduced by normalising the SNP density by length. The absolute number of homozygous SNPs in each contig is divided by the number of nucleotides (contig length) so we obtain the contig score:

$$Score_n = \frac{\sum Hom}{length_n}$$

SDM sorts the contigs based on their score so that they follow an ideal normal distribution. It starts by taking the 2 lowest values that should be at both tails of the distribution. Following this fashion, we obtained the right and left sides that together build up the whole distribution.

The first SDM version was run on the model genomes created as explained in the previous section. SDM uses the VCF file with the homozygous and heterozygous SNP positions, the text files containing the lits of homozygous and heterozygous SNPs and the FASTA file with the shuffled contigs as input. The FASTA file with the correct contig ordered is used to calculate the ratios in the correctly ordered fragments so that they can be compared to the ratios obtained after SDM sorts the contigs. The command lines to run SDM on the model genomes are available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Small_genomes/SDM.sh.

SDM generates a new FASTA file with the suggested contig order, plots comparing the hypothetical SNP densities to the expected densities, a plot comparing the real ratio distribution to the ratio distribution after running SD and a CSV

For all the 70 genomes ranging from 1 to 15 Mb, no filtering step based on the ratio was used (threshold = 0). The highest kernel density value for the SNP distribution after sorting the contigs with SDM was taken as candidate SNP. Since the peak of the SNP homozygous distribution (mean) was known, the peak obtained after SDM was compared to the original value to measure the deviation of the approach:

$$Deviation = \frac{|Candidate - Causative|}{Length}$$

where ‘Candidate’ is the SDM predicted position and ‘Causative’ is the mean of the normal distribution of homozygous SNPs in the model genome. A CSV file containing all the deviations in the model genomes was built and it’s available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Small_genomes/arabidopsis_datasets/1-15Mb.csv. The same approach was used for the whole-sized genomes (chr1_i, chr1_A_i and chr1_B_i).

The Ruby code used to run SDM on model genomes is available in the Github repository https://github.com/pilarcormo/SNP_distribution_method/blob/master/

Small_genomes/SNP_distribution_method.rb.

2.3. Jitter plots for SDM deviation

The deviation percentages calculated independently for each genome as explained in section 2.2 are available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Small_genomes/1-15Mb.csv and https://github.com/pilarcormo/SNP_distribution_method/blob/master/Small_genomes/30Mb.csv. The R code to plot the deviation jitter plots for each genome length and contig size was deposited at https://github.com/pilarcormo/SNP_distribution_method/blob/master/R_cripts/jitter_plots.R

2.4. Pre-filtering step based on ratio

The homozygous to heterozygous SNP ratio was used as a cut-off value to discard contigs located further away from the causal mutation. If this filtering step is required, the threshold astringency should be provided as an integer (1, 5, 10, 20). Each integer represents the percentage of the maximum ratio below which a contig will be discarded. In example, if 1 is specified, SDM will discard those contigs with a ratio falling below 1% of the maximum ratio while a value of 20 is more astringent will discard those contigs with a ratio falling below 20% of the maximum ratio.

We used the model genomes defined on section 2.1 to test the effectiveness of the filtering step. In particular, we used the replicates of the genome with the normal distribution peak shifted to the right and the replicates of the genome with the normal distribution peak shifted to the left (chr1_C_i and chr1_E_i, respectively). Protocol and results were deposited at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Small_genomes/arabidopsis_datasets/Analyse_effect_ratio. The folders /chr1_right and /chr1_left contain examples of the SDM output after filtering under the names Ratio_0_1 (no filtering), Ratio_1_1 (1% threshold), Ratio_5_1 (5% threshold), Ratio_10_1 (10% threshold), Ratio_20_1 (20% threshold).

2.5. Forward genetic screens used to analyse SNP distribution

We used five different sets of Illumina sequence reads from 4 recent out-cross [@Galvão et al. 2012], [@Uchida et al. 2014] and back-cross experiments [@Allen et al. 2013], [@Monaghan et al. 2014] in *Arabidopsis thaliana* backgrounds **(table 1)**.

Galvão et al obtained the first set of reads (**OCF2**) by sequencing a mutant pool of 119 F2 mutants generated by out-crossing a Col-0 background mutant to a Ler-0 mapping line. They also sequenced the parental lines and performed conventional SHOREmap [@Schneeberger:2009] to identify the mutation [@Galvao:2012]. The reads are available to download at http://bioinfo.mpipz.mpg.de/shoremap/examples.html

In the second study (**BCF2** dataset), Allen et al back-crossed the Col-0 mutant to the non-mutagenized Col-0 parental line [@Allen et al. 2013]. A pool of 110 mutant individuals showing the mutant phenotype and the parental line were sequenced. They used different SNP identification methods that produced highly similar outcomes (NGM, SHOREmap, GATK and samtools) [@Austin:2011], [@Schneeberger:2009], [@DePristo:2011aa], [@Li:2009aa]. The reads are available to download at http://bioinfo.mpipz.mpg.de/shoremap/examples.html

The third study we analysed was also a back-cross experiment. Monaghan et al obtained two different and independent Col-0 mutants (called **bak1-5 mob1** and **bak1-5 mob2**) [@Monaghan:2014]. The mutants were back-crossed to a parental Col-0 line and sequenced. They used CandiSNP [@Etherington:2014] to identify the causal mutation.

The last dataset we used (**sup#1** dataset) was obtained by outcrossing a Arabidopsis Wassilewskija (Ws) mutant to wild-type Col-0 plants followed by sequecing of 88 F2 individuals and Ws and Col as parental lines. Uchida et al described a pipeline to identify the causal mutation based on the peaks obtained by plotting ratios of homozygous SNPs to heterozygous SNPs [@Uchida et al. 2014]. Reads are available at http://www.ncbi.nlm.nih.gov/sra/?term=DRA000344

2.6. Read mapping and SNP calling

Mutant and parental reads were subjected to the same variant calling approach. The Rakefile and scripts used to perfom the alignment and SNP calling can be found in the Suplementary file 1

The quality of the deep sequencing was evaluated using FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed and quality filtered by Trimmomatic v0.33 [@Bolger:2014aa]. We performed a sliding window trimming, cutting once the average Phred quality fell below 20 in the window.

The paired-end reads were aligned to the reference sequence of *Arabidopsis thaliana* TAIR10_chr_all.fas at ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_chromosome_files/ [@Lamesch:2012aa] from The Arabidopsis Information Resource by BWA-MEM long-read alignment using BWA v 0.7.5a with default settings [@Li:2010]. The resultant alignment (SAM files) were converted to BAM file and then sorted using the samtools package v1.0. Then, we used samtools mpileup command to convert the BAM files into pileup files. To call SNPs we used the mpileup2snp command from VarScan v2.3.7 http://varscan.sourceforge.net [@Koboldt:2012aa], [@Koboldt:2009aa] to get VCF 4.1 output. A default allele frequency of 0.8 was used to clasify the SNPs as homozygous.

VCF files for mutants and mapping lines can be found at the repository https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads in the individual folder for each screen (/OCF2, /BCF2, /Aw_sup1-2, /m_mutants)

for OCF2 mutant is at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/OCF2/OF_output25vcf.zip and the parental VCF is at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/OCF2/Ler/OC_parent.vcf.zip.

The whole pipeline used for readn mapping, SNP calling is summarised in Additional figure 1

2.7. Parental filtering

To improve SDM accuracy and unmask the high homozygous SNP peak, we performed a filtering step to reduce the SNP density in the mutant VCF files based on the background SNPs. The parental reads were also mapped to the Arabidopsis reference genome as explained in section 2.6 followed by a step of SNP calling. The SNPs present in the non-mutant parental reads were obviously not induced by the mutagen (EMS) and can be considered as ‘background’ mutations (not responsible for the mutant phenotype).

The workflow used to filtered the reads can be found in the Suplementary file 2 and the protocol is available in the README file deposited at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Reads.

The code used is available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/manage_vcf.rb

2.8. Centromere removal

A great part of the variability observed in the genomes was due to the presence of centromeres. The code at https://github.com/pilarcormo/SNP_distribution_method/blob/master/remove_cent.rb was used to discard the SNP positions that were due to the centremere variability. The workflow used to filtered the reads can be found in the Suplementary file 2 and the how-to is available in the README file deposited at https://github.com/pilarcormo/SNP_distribution_method/tree/master/Reads

2.9. SNP density analysis

The absolute number of homozygous SNPs before and after filtering was taken from the reads folder by running SNP_density.rb. The command lines are available at Suplementary file 2.

The output CSV file is available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/density.csv. It shows the number of homozygous SNPs per chromosome and pwe forward genetic screen (BCF2, OCF2, sup#1, mob1, mob2). After obtaining the total number of homozygous SNPs per genome by adding together the values per chromosome, we wrote new CSV files for the back-cross and the out-cross experiments. Those are available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/density_sum_back.csv and https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/density_sum_out.csv.

We used the R code at https://github.com/pilarcormo/SNP_distribution_method/blob/master/R_scripts/SNP_filtering.R to plot the total number of homozygous SNPs before filtering, after parental filtering and after centromere removal.

Then, the homozygous and heterozygous SNP densities obtained after filtering for each study were plotted together with the ratio signal to identify the high density peaks in the distribution. The R code was deposited at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/filtering.md

2.10. Probability plots

To analyse the correlation of the homozygous SNP density in forward genetic screens to a normal distribution, probability plots (QQ-plots) were created. We used the homozygous SNP positions in the chromosome were the causative mutation was located after parental filtering and centromere removal. The R code and plots are available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/qqplot.md

2.11. Kurtosis

The R code and a table showing the results obtained after analysing the shape and variance of the normal distributions of homozygous SNPs is available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Reads/kurtosis.md

2.12. Analysis of average contig size in different whole genome assemblies

Genome assemblies at contig level available for plants at http://www.ncbi.nlm.nih.gov/assembly/organism/3193/all/ were used to define a more realistic N50 contig size in our model genomes. All the contig assemblies from January 2013 until June 2015 which provided a full genome representation with a genome coverage higher than 1x were analysed. Only those providing the sequencing technology and the N50 contig size were selected to analyse the contig size distribution.

The whole table with the chosen assemblies and the results are available at this repository https://github.com/pilarcormo/SNP_distribution_method/tree/master/Contigs.

We plot the N50 contig size against genome size using a different colour for each sequencing technology. Then, we calculated the N50 density and the median of the distribution. We focused on the 16 assemblies built with Illumina Hiseq and we tried tried to define a model for the N50 contig size change over genome length. We applied a Generalized Additive Model (GAM) to fit non-parametric smoothers to the data without specifing a particular model. First, we applied logarithms to both N50 size and genome length on the Illumina HiSeq assemblies. The R code can be found at https://github.com/pilarcormo/SNP_distribution_method/blob/master/Contigs/contigs.R

2.13. Model genomes based on real SNP densities

We created new model genomes using the homozygous and heterozygous SNP densities obtained from the forward genetic screens after parental filtering and centromere removal. Those files were. Three minimum contig sizes were used (2,000, 5,000 and 10,000 bp), being the maximum values 4,000, 10,000 and 20,000 bp respectively. The contig sizes oscillated between the minimum and the maximum values. Instead of using idealised SNP distributions as explained in section 2.1, we used the homozygous and heterozygous SNP lists after parental filtering and centromere removal. The densities were deposited at https://github.com/pilarcormo/SNP_distribution_method/tree/master/arabidopsis_datasets/SNP_densities.

The genomes were generated by running https://github.com/pilarcormo/SNP_distribution_method/blob/master/model_genome_real_hpc.rb. The command lines are available at Suplementary file 2

The genomes are available at https://github.com/pilarcormo/SNP_distribution_method/tree/master/arabidopsis_datasets/No_centromere. They are classified by contig size. The directories contain the following: frags_shuffled.fasta, frags.fasta, hm_snps.txt, ht_snps.txtand snps.vcf

2.14. SDM with real SNP densities

The model genomes generated as explained in 2.13 were used to prove the SDM efficiency to identify the genomic region carrying the causative mutation. The Ruby code is available at https://github.com/pilarcormo/SNP_distribution_method/blob/master/SNP_distribution_method_variation.rb. The input and output specification for SDM can be found at the README file in the main project Github repository https://github.com/pilarcormo/SNP_distribution_method.

Instead of specifying a percentage of the maximum ratio to filter the contigs, we used an automatic approach that tailor the threshold for each specific SNP density anc contig length. The default percentage of the maximum ratio used was 1%. After the first filtering round, if the number of dicarded contigs is below a 3% of the starting number of contigs, the percentage of filtering is increased by 2 (it will be 2% in the first repetition step) and the filtering is repeated until the condition specified is met.

The command lines used to run SDM on the model genomes generated as explained in section 2.13 are available at Suplementary file 2.

Results.md 3. Results and discussion ===

3.1. SDM works over a range of effective genome lengths and realistic fragment sizes

We created model genomes based on *Arabidopsis thaliana* chromosomes to develop our mutant identification method. We chose *Arabidopsis thaliana* to create our model genomes because it is been a widely used organism for forward genetic screening and it is a relatively small and well-annotated genome. Also, SNP densities in several mapping-by-sequencing experiments in Arabidopsis are available (see section 3.3 for several examples) so they could be used as a basis to develop our methodology.

By making customised genomes, we could rapidly alter different paramenters as genome length, contig size or SNP density to analyse their effect on the detection method accuracy. Our dinamic way of creating model genomes helped us define all the different aspects that should be taken into account when analysing the SNP distribution. Also, the causative mutation was defined manually by us, so we could measure the deviation between defined and predicted value. We generated the model genomes by asigning an idealised SNP distribution to a set of randomly shuffled sequences that imitate contigs assembled from HTS. In the model genomes, homozygous SNPs follow a normal distribution while heterozygous SNPs follow a uniform distribution.

To estimate the effect of the genome length on SDM's ability to find the causative mutation, we created model genomes of different sizes ranging from 1 Mb to 15 Mb. We defined a variable contig length randomly ranging between a provided mininum value and its double. We used two minimum contig sizes (500 bp and 1000 bp) that gave rise to approximately 1300 and 700 contigs respectively. To guarantee the confidence of SDM, we created 5 replicates for each condition. For the second set of model genomes, we used the whole chromosome I length and a more realistic SNP density (1 SNP every 3000 bp). In this case, 3 contig sizes were employed (1000, 2000 and 4000 approximately).

The SNP Distribution Method (SDM) sorts the sequence fragments by their SNP density values (the score calculated as defined in section 2.2) so that they follow a normal distribution. Then, a second sorting step in which the contigs are also sorted by their ratio (calculated as defined in section 2.2) is applied. We considered the highest kernel density value for the SNP distribution after SDM sorting as the candidate SNP location and this value was compared to the peak of the normal distribution previously defined in the model genome. The difference was called 'deviation' and can be found in **Fig.2**. Consistent results were obtained for all the replicates. SDM identified the high density peak with no significant effect of genome length and contig size. The deviation from the causative mutation assigned in the model was lower than 1% in the small SNP-riched genomes (Fig 2A). This was also true for the whole-sized genomes when they were divided in 1000 and 2000 contigs but not for the 4000 contigs (Fig 2B), where 4 replicates had a deviation between 1 and 3%. Even though the deviation is still low, the higher the number of contigs, the harder it is to get to the correct order. Also, as we did not change the SNP density, the SNPs are spreaded over more fragments and the sorting is more complicated.

3.2. A pre-filtering step based on the homozygous to heterozygous SNPs ratio improves SDM accuracy

As explained in the previous section, SDM succeed in the high density peak identification in model genomes when the idealised causative mutation was located in the middle of the distribution as the number of fragments at both sides (right and left) was the same. However, this success is only true in the high SNP density area (peak of the distribution) and when we shifted the causal mutation in our model to one side (one tail was longer than the other), SDM was not able to sort the contigs in the tails properly. Even though the contigs located in the peak were those in which the mutation was defined, the algorithm was not able to classify the low density contigs as belonging to the right or left tail and the distribution peak appeared moved, giving rise to a high deviation from the defined peak.

To fix the problem, we delimited a threshold value based on the ratio of homozygous to heterozygous SNPs to discard contigs located further away from the causative mutation. We excluded those contigs with a ratio below a given percentage (1%, 5%, 10% or 20%) of the maximum ratio. Therefore, only those contigs with a high ratio are sorted. Even though the exact position in the genome cannot be determined by this approach, we can assess the contigs in which the mutation is more likely to be found, dismissing a great part of the genome that is hiding the causative mutation. To test the influence of the threshold we used the whole chromosome I from *Arabdidopsis* as a model genome (**Fig. 3**). The high density peak in the ratio distribution matched the expected when aproximately a

80-90% of the genome is discarded. In this case, the SNP density was high (1 SNP every 3000 bp) and the distribution peak was narrow. However, we cannot establish a standard threshold, since it will depend on the peak deviation on each case. To avoid lossing interesting mutations by being too strict with the filtering, we suggest using 1-5% of the maximum ratio to obtain optimal results.

3.3. Filtering background SNPs and centromeres unmasks the high homozygous SNP density peak in several bulk segregant analysis in Arabidopsis

We selected different sets of data of bulk segregant analysis of a mutation segregating in an out-cross [@Galvao:2012], [@Uchida:2014] or back-cross [@Allen:2013], [@Monaghan:2014] population. We performed conventional genome alignment to the reference genome using the reads provided in the 4 studies. The techniques used to identify the mutations (**Table 1**) were different in every case so we took advantage of this fact to confirm the reproducibility of the available methodology to identify causative mutations. We aligned the Illumina paired-end reads to the *Arabidopsis thaliana* reference genome. Then, we used VarScan for variant calling.

In the first out-cross experiment we analysed (OCF2), Galvão et al identified the mutation causing late flowering which lied on chromosome 2, specifically on gene *SOC1* (18807538..18811047) **[@Galvao. 2012]** . In the second study (BCF2) Allen at al analised the mutant individuals showing leaf hyponasty to identify a gene involved in the Arabidopsis microRNA pathway [@Allen et al. 2013]. They identified the causal SNP mutation in *HASTY*, a gene on chromosome 3 (1401271..1408197). Also, we used the reads from a forward screen done in the immune-deficient bak1-5 background aim identify new components involved in plant immunity. Monaghan et al found 2 causative mutations in the gene encoding the calcium-dependent protein kinase CPK28 (26456285..26459631) for both bak1-5 mob1 and bak1-5 mob2 [@Monaghan:2014]. In the last mapping-by-sequencing study we used, Uchida et al reported that the majority of SNPs detected on chromosome 4 were homozygous and they identified the sup#1 mutation on the *SGT1b* gene (6851277..6853860) [@Uchida:2014] (**Table 2**)

The main advantage of working with already identified mutations was to focus directly on the chromosome were the mutation was previously described. We analysed the total number of homozygous SNPs in the chromosome where the causative mutation was located. As expected, when the mutant individual is out-crossed to a distant mapping line (OCF2 and sup#1), the SNP density is up to 20 times higher than in the case of back-crossing to the parental line (BCF2 and mob mutants). In the back-crossed populations, we identified approximately 1700 homozygous SNPs. That gives an overall density of 1 SNP every 15000 bp. We identified 9200 homozygoys SNPs in the chromosome of interest in the first out-crossed population (OCF2) and 27000 SNPs in the second out-crossed population (sup#1). The overall density was of 1 SNP every 2500 bp for OCF2 and 1 SNP every 700 bp for sup#1.

Parental filtering was fundamental to reduce the SNP density and unmask the SNP linkage around the causative mutation and it was especially crucial in out-crossed populations where the starting density was higher. The parental lines used to back or out-cross were also sequenced and mapped to the Arabidopsis thaliana reference genome. The SNPs present in the non-mutant parental reads were onsidered as 'background' mutations and filtered from the mutant SNP lists. In the back-cross studies, the absolute homozygous SNP number was reduced up to 9 times (**Fig4A**) after parental filtering. The total number of homozygous SNPs was reduced 3 times in out-crossed populations (**Fig4C**). Even though the centromere removal did not reduce the total number of SNPs in the same proportion as parental filtering did, it was essential to unmask the normal distribution around the causative mutation. Many studies have shown the centromeres peculiarity, characterised by high repeat abundance (often >10,000 copies per chromosome) [@Melters:2013aa]. This high variability in a few hundred of bp generates a high SNP density peak which hides the peak of interest.

The filtering steps reduced the complexity of the distribution (Additional fig 3) and improved the degree of correlation to a normal distribution (see section 3.5). The usefulness of removing non-unique SNPs is not new, and all the mapping-by-sequencing experiments we used did the same filtering to some extent. ([@Galvao:2012], [@Uchida:2014]], [@Allen:2013], [@Monaghan:2014])

We identified a unique peak in the area where the causative mutation was described when we plotted the homozygous SNP density obtained after filtering. We calculated the homozygous to heterozygous ratio for each contig and the ratio values were overlapped to the SNP densities, obtaining the distributions in **Fig 5**. All the densities were congruent with the results described in the publications.

3.4. Homozygous SNPs in forward genetics screens are normally distributed around the causative mutation

After running the variant calling approach with the different datasets, we showed a unique peak in the SNP distribution around the causative mutation (**Fig. 3**). The next step was to analyse the correlation of the SNP distribution to a theoretical probability distribution. We created probability plots or Q-Q plots with the homozygous SNP densities in the back-crossed and out-crossed populations (**Fig 6**). Our results indicate a good correlation between the homozygous SNP frequencies and a normal distribution. We further validated the correlation between the sample values and the predicted values by a simple linear regression (r2 >0.9). The standard deviation, although variable between samples, oscillated between 3 and 7 Mb (**Table 2**). Therefore, this is evidence that 15 Mb model genomes are large enough to identify the normal density peak when using SNP densities from real forward genetic screens.

We also analysed the shape of the distributions by measuring the kurtosis (**Table 3**). 4 of the 5 distributions were platykurtic, as they showed a negative kurtosis and consequently a lower, wider peak around the mean and thinner tails. Only one (sup#1) was leptokurtic (positive kurtosis). This might be due to the high density observed in sup#1 (more than 3 times higher than the other out-crossed population) which generates a narrower peak due to the greater linkage of SNPs in the non-recombinant area. In the other examples, the number of SNPs linked together is lower and they are highly scattered in the genome, generating the lower and wider peak.

3.5. The N50 contig size in plant genome assemblies depends on genome size and sequencing technology

We planed to generate more realistic model genomes. To decide on a representational contig size, 29 assemblies at contig level were analysed. The relationship between genome length and N50 contig size was not really obvious as other aspects as the sequencing technology used **(Fig7A)** or the genome coverage have a high impact on the final N50 contig size. We observed that the preferred sequencing technology in the last 2-3 years is Illumina HiSeq since half of the assemblies were built using this technology. Other assemblies used combinations of different techniques (including Illumina, 454 and PacBio). The median value of the N50 contig length for all the 29 assemblies is 11,517 bp while it is reduced to 5,484 bp when analysing only HiSeq assemblies (**Fig7B**). To decrease the effect of the technology used, we focused on the 16 assemblies built with Illumina Hiseq and we tried to establish a model that could explain the N50 size change over genome size.

The relationship was not linear but we did not have any mechanistic model to describe it, so we apply a Generalized Additive Model (GAM) to fit non-parametric smoothers to the data without specifing a particular model. First, we applied logarithms to both N50 size and genome length on the Illumina HiSeq assemblies. When we focused on those genomes larger than 200 Mb, the R square was 0.807 and p-value was 0.000909.

We used 3 different contig sizes to create the model genomes. The first two model genomes were built using the N50 median values. We chose 5,000 bp (based on the median for Illumina Hiseq) and 10,000 bp (based on the median for all the assemblies). The last contig size was decided looking at the model defined for Illumina HiSeq. Due to the *Arabidopsis thaliana* genome size, the minimum contig size decided for these model genomes was 2,000 bp.

3.6. SDM identifies the genomic region carrying the causal mutation previously described by other methods

As a final proof-of-concept, we used the SNP densities obtained from OCF2, BCF2, mob1, mob2 and sup#1 datasets as explained in section 3.3 after parental filtering and centromere removal to build new model genomes. We split the *Arabidopsis thaliana* chromosomes where the mutations into fragments as described in section 3.5.

We regained the normal distribution for all the datasets after shuffling the contig order and running SDM. We created a plot to compare the hypothetical ratio distribution after SDM and the real one to test the performance of SDM. The results for all the model genomes generated were deposited at a Github repository at https://github.com/pilarcormo/SNP_distribution_method/tree/master/arabidopsis_datasets/No_centromere.

Our prior knowledge about the correct contig order allowed us to define the real chromosomic positions in the artificially-made contigs identified by SDM as candidates. In that way, we were able to adjust the method to get its best efficiency

(Table 4).

Contig size had an effect on the number of candidate contigs supplied by SDM. When the minimum contig sizes were 2 and 5 kb, the SNP positions were split to different contigs and it was more difficult for SDM to find the correct contig order. When the average contig size was below 10 kb, 20 candidate contigs were considered for back-crossed populations and 40 were needed for out-crossed populations due to the high SNP densities. However, when the average contig size was over 10 kb, 12 candidate contigs in the middle of the distribution were enough to contain the causative mutation. The candidate region ranges from 60 to 180 kb depending on the contig size and the type of cross.

In our example, the number of candidate positions was higher for the out-crosses than for the back-crosses. This is logical as the SNP density after filtering was higher in the out-crossed populations, and therefore, the number of segerating SNPs in the candidate contigs is larger.

We could not define a universal cut-off value based on the homozygous to heterozygous ratio for all the different SNP densities, as sometimes the region with a high ratio was narrow due to a high SNP linkage in the area, while in other cases, the increase in the ratio was progressive, and the peak was wider. Therefore, when we worked with real densities, we used an automatic approach that tailor the threshold for each specific SNP density anc contig length. Table 4 shows the tailored thresholds and total discarded contigs for all the datsets.

We can conclude that SDM is a rapid and precise method to perform bulk segregant linkage analysis from back-crossed and out-crossed populations without relying on the disponibility of a reference genome, specially effective on large contig sizes/

Conclusions

Forward genetic screens are very useful to identify genes responsible for particular phenotypes. Thanks to the advances in HTS technologies, mutant genomes sequencing has become quick and unexpensive. However, the mapping-by-sequencing methods available present certain limitations, complicating the mutation identification especially in non-sequenced species. To target this problem, we proposed a fast, reference genome independent method to identify causative mutations. We showed that homozygous SNPs are normally distributed in the mutant genome of back-cross and out-crossed individuals. Based on that idea, we defined a theoretical SNP distribution used by SDM to identify the genomic region where the causative mutation was located. We conclude that SDM is especially sucessful for analysing mutants obtained from a back-cross population. The increase in the number of SNPs in out-cross experiments complicated the genetic analysis and the mutation estimation. Ideally, over the next few years, sequencing costs will decrease and this will allow to sequence every mutant individual from a forward genetic screen. Therefore, we need fast and reliable methods to identify variants bypassing the reference genome assembly step. We now aim to improve and apply SDM in forward genetics screens of species where a reference genome is not yet available. We plan to develop an accessible software that will speed up gene finding in non-sequenced organisms.

Tables.md **Table 1**. Forward genetics screens developed in *Arabidopsis thaliana*, the lines involved in the crossing and the technologies used to identify the causative mutation in each case.

Study	Sample	Mutant		Wild-type	SNP caller
Galvão et al. 2012	OCF2	Col-0	x	Ler-0	SHOREmap
Allen et al. 2013	BCF2	Col-0	x	Col-0	NGM, SHOREmap, GATK and SAMtools
Monaghan et al. 2014	bak1-5 mob1 and mob2	Col-0	x	Col-0	CandiSNP
Uchida et al. 2014	sup#1	Ws-0	x	Col-0	Ratio of homozygous to heterozygous SNPs

Table 2. Out-crossed and back-crossed populations, chromosome where the mutation was found and mutated gene location.

Sample	Cross	Chr	Mutated gene
OCF2	Out-cross	2	SOC1(~18.8 Mb)
BCF2	Back-cross	3	HASTY (~14.05 Mb)
mob1/mob2	Back-cross	5	CPK28 (~26.45 Mb)

Sample	Cross	Chr	Mutated gene
sup#1	Out-cross	4	SGT1b (~6.85 Mb)

Table 3. Measurement of the homozygous SNP density correlation to a theoretical normal distribution in several out-cross and back-cross experiments. Analysis of variance and distribution shape (kurtosis and skewness).

Sample	SNPs	Chr	r2	SD (Mb)	Kurtosis	Skewness
OCF2	151	2	0.949	6.01	1.85	-0.392
BCF2	15	3	0.959	3.20	2.02	0.201
mob1	25	5	0.944	7.20	2.69	-0.240
mob2	41	5	0.894	3.71	2.37	0.445
sup#1	4633	4	0.976	3.66	3.50	0.370

Table 4. SDM mutant identification success when using an automatic filtering approach to discard contigs. 3 different contig sizes analysed, the percentages of the maximum ratio used as threshold are specified. In brakets, the number of contigs discarded out of the total number of contigs.

Sample	Cross	Chr	Contig size (kb)	Threshold	Identification
OCF2	Out-cross	2	2-4	5% (230/6568)	**Unsucessful**
			5-10	3% (189/2634)	Sucessful
			10-20	3% (186/1328)	Sucessful
BCF2	Back-cross	3	2-4	35% (7807/7821)	Sucessful
			5-10	21% (108/3130)	Sucessful
			10-20	21% (95/1562)	Sucessful
mob1	Back-cross	5	2-4	17% (254/8992)	**Unsucessful**
			5-10	15% (239/3603)	Sucessful
			10-20	15% (220/1805)	Sucessful
mob1	Back-cross	5	2-4	35% (8950/8994)	Sucessful
			5-10	21% (195/3582)	Sucessful
			10-20	21% (189/1804)	Sucessful

Sample	Cross	Chr	Contig size (kb)	Threshold	Identification
sup#1	Out-cross	4	2-4	3% (228/6201)	Sucessful
			5-10	3% (153/2491)	Sucessful
			10-20	3% (93/1240)	Sucessful

figures.md

figures_descriptions.md **Figure 1.** SDM workflow.

Figure 2. SDM percentage of deviation from the causative mutation expected location in the model genomes. The deviation is a measure of the difference between the expected mutation position and the candidate position predicted by SDM normalised by the model genome length. 5 replicates of each model genome were created. **(A)** Genome size range from 1 to 15 Mb with two contig sizes (700 and 1300 contigs). **(B)** Whole-sized *Arabidopsis thaliana* chromosome 1 with three contig sizes (4000, 2000 and 1000 bp)

Figure 3. Relevance of a pre-filtering step based on the homozygous to heterozygous SNP ratio in model genomes. The ratio is calculated by contig and those contigs falling below a given percentage of the maximum ratio are discarded. The expected ratio was measured in the corretly ordered fragments. Thge SDM ratio was measured after SDM sorting. **(A)** No filtering. **(B)** 1% of the maximum ratio **(C)** 5% of the maximum ratio **(D)** 10% of the maximum ratio

Figure 4.. Absolute number of homozygous SNPs before and after filtering in independent **(A)** back-crossed and **(B)** out-crossed populations. The final candidate positions after running SDM were also compared in the **(C)** back-crossed and **(D)** out-crossed populations.

Figure 5. Identification of high homozygous SNP density peaks surrounding the causal mutation in 5 independent studies. **(A)** Overlapping homozygous, heterozygous SNP densities and ratios for OCF2, BCF2, bak1-5 mob1/mob2 and sup#1.

Figure 6. Measurement of the homozygous SNP density correlation to a normal distribution in back-crossed and out-crossed populations by probability (Q-Q) plots. Simple linear regression was used to determine the relationship and the

Figure 7. Analysis of average contig size in different whole genome assemblie. **(A)** N50 contig size vs Genome size in 29 whole genome assemblies at contig level. Colour represents the sequencing technology or the combination of sequencing technologies used. **(B)** N50 contig size distribution for Illumina HiSeq (pink) and the

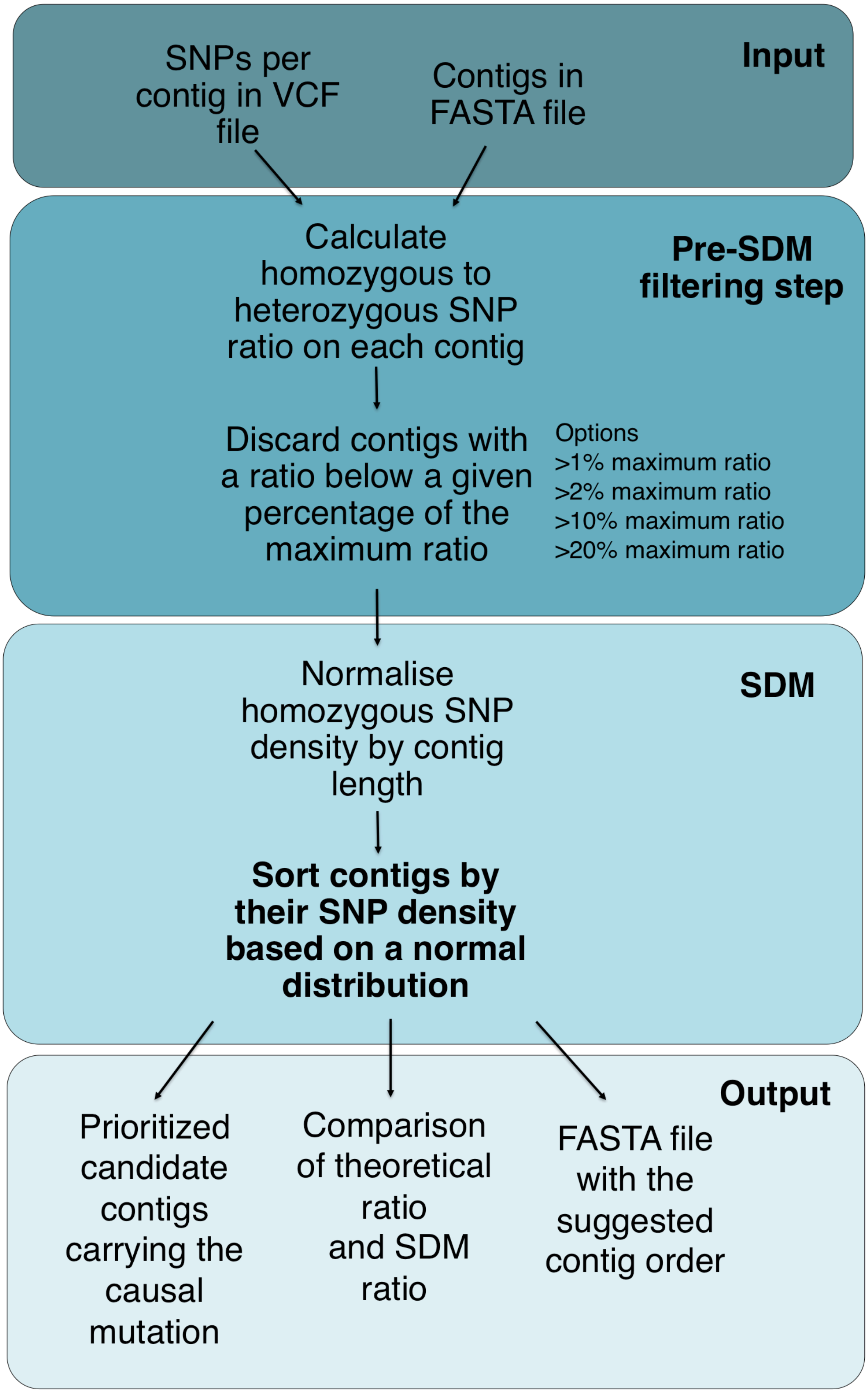
Input
SNPs per contig in VCF file
Contigs in FASTA file
Pre-SDM filtering step
Calculate homozygous to heterozygous SNP ratio on each contig
Discard contigs with a ratio below a given percentage of the maximum ratio
Options
>1% maximum ratio
>2% maximum ratio
>10% maximum ratio
>20% maximum ratio
SDM
Normalise homozygous SNP density by contig length
Sort contigs by their SNP density based on a normal distribution
Output
Prioritized candidate contigs carrying the causal mutation
Comparison of theoretical ratio and SDM ratio
FASTA file with the suggested contig order

Figure 1

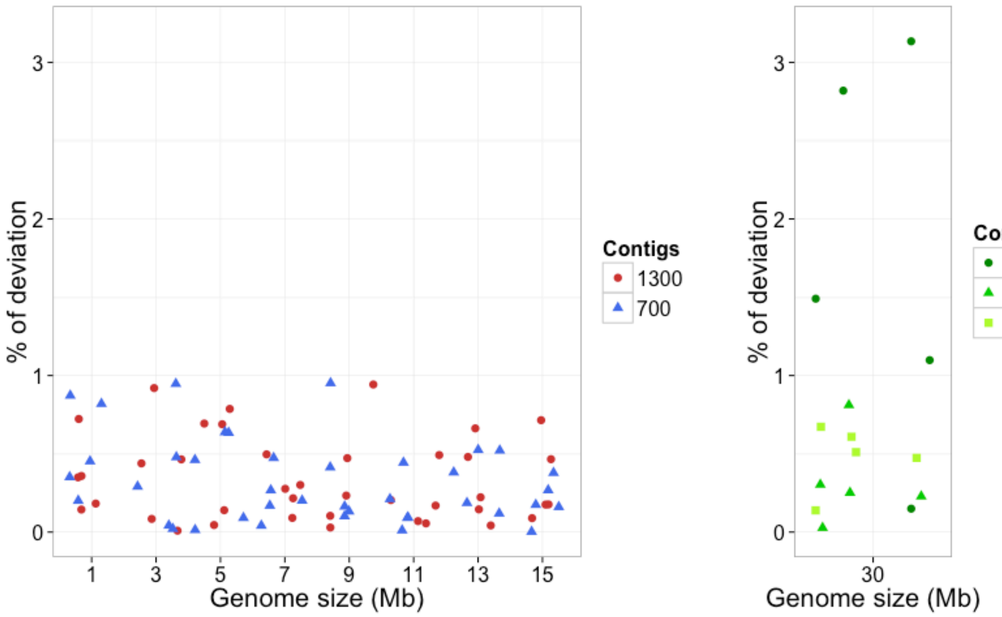

Figure 2

other sequecing technologies (blue). Medians are represented by the dashed lines. **(C)** Model for the non-linear relationship between N50 contig size and genome size in Illumina HiSeq assemblies

Additional_fig_descriptions.md #Additional figures

Additional figure 1 Standard pipeline for sequence alignment and SNP calling in forward genetics screens in *Arabipdopsis thaliana*

Additional figure 2 Differences in the Hom/het ratio density by adding a different factor (1, 0.1 and 0.0.1) to numerator and denominator.

Additional figure 3 Change in the homozygous SNP density plot before filtering, after background SNPs filtering and centromere removal. Additional_figures.md

sup_file1.md Supplementary file 1. Method workflow ===

Quality filtering for paired-end reads

```
FastQC/fastqc reads_R1.fq
FastQC/fastqc reads_R2.fq
```

```
java -jar Trimmomatic-0.33/trimmomatic-0.33.jar PE reads_R1.fq reads_R2.fq paired_R1
```

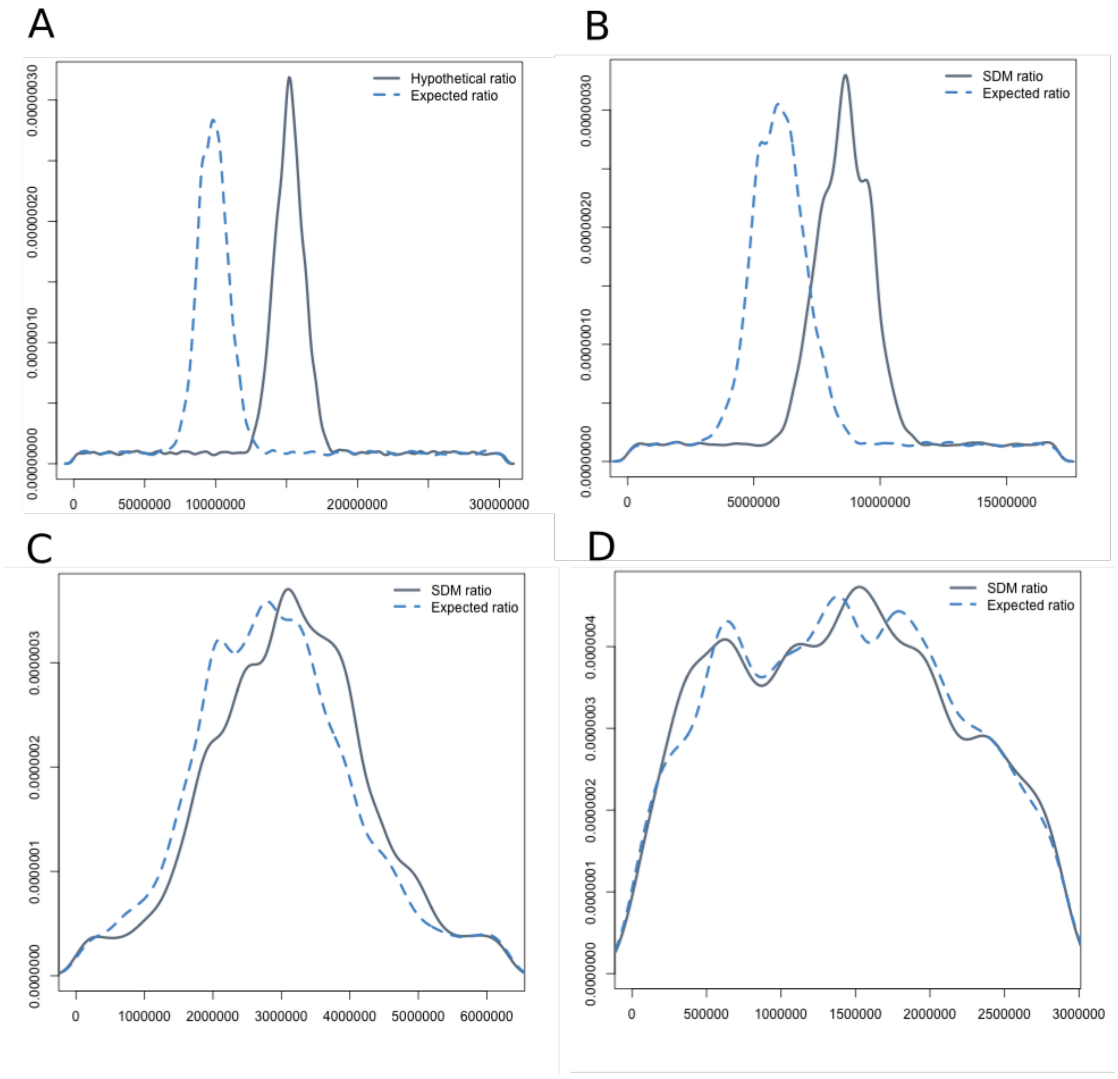
A
Hypothetical ratio
Expected ratio
0.00000000
0.00000010
0.00000020
0.00000030
0
5000000
10000000
20000000
30000000
B
SDM ratio
Expected ratio
0.00000000
0.00000010
0.00000020
0.00000030
0
5000000
10000000
15000000
C
SDM ratio
Expected ratio
0.0000000
0.0000001
0.0000002
0.0000003
0
1000000
2000000
3000000
4000000
5000000
6000000
D
SDM ratio
Expected ratio
0.0000000
0.0000001
0.0000002
0.0000003
0.0000004
0
500000
1000000
1500000
2000000
2500000
3000000

Figure 3

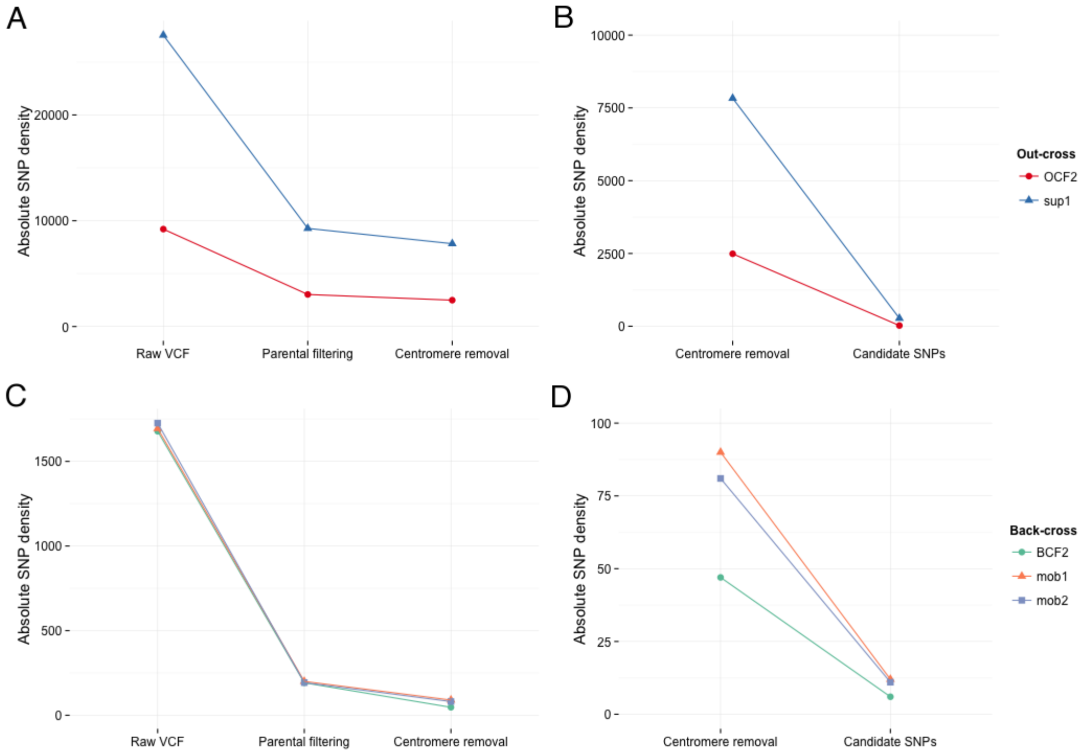

Figure 4

Command line parameters used for paired-end read mapping and SNP calling

1. Index reference sequence

```
bwa index TAIR10.fa
```

2. Map the reads to reference genome with BWA

```
desc "Align using bwa"
task :bwa  do
     sh 'bwa mem TAIR10.fa paired_R1.fq paired_R2.fq > alignment.sam'
end
```

3. Convert the respective SAM file to BAM file and sort the BAM file using samtools

```
desc "Convert sam to bam file"
task :bam => ["bwa"] do
```

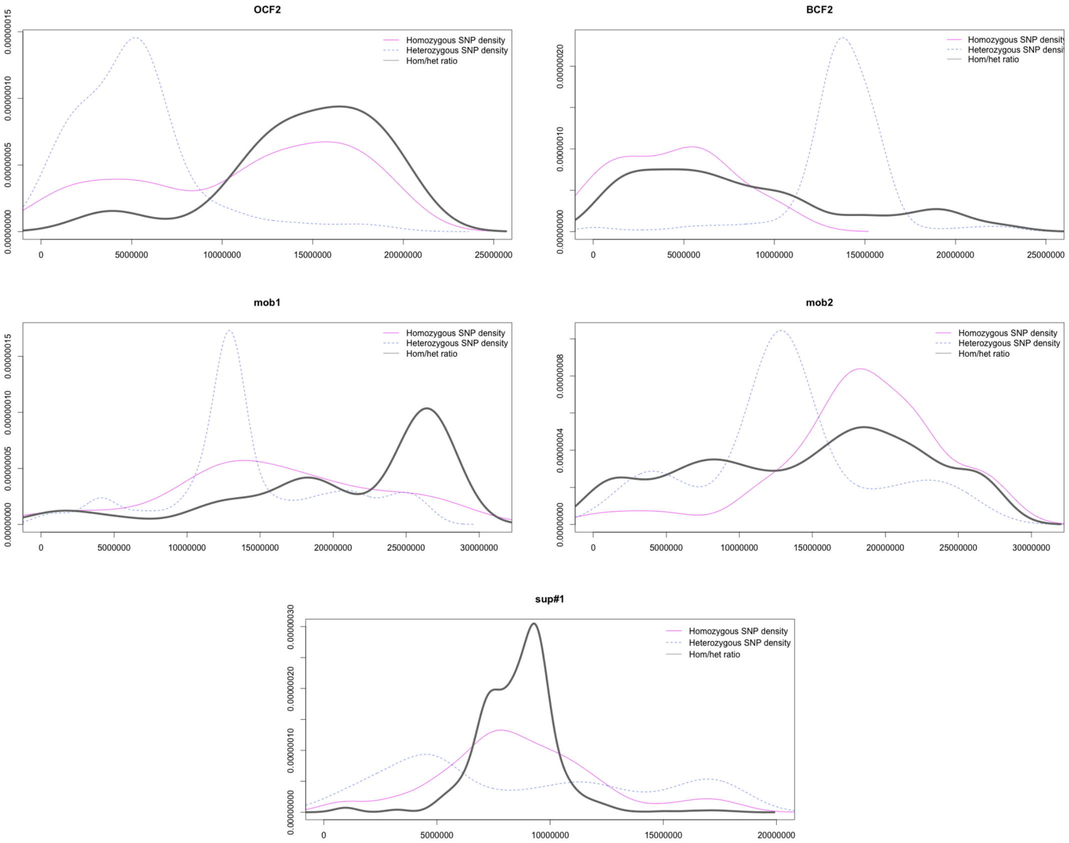
OCF2
Homozygous SNP density
Heterozygous SNP density
Hom/het ratio
0.00000000
0.00000005
0.00000010
0.00000015
0
5000000
10000000
15000000
20000000
25000000
BCF2
Homozygous SNP densit
Heterozygous SNP densi
Hom/het ratio
0.00000000
0.00000010
0.00000020
0
5000000
10000000
15000000
20000000
25000000
mob1
Homozygous SNP density
Heterozygous SNP density
Hom/het ratio
0.00000000
0.00000005
0.00000010
0.00000015
0
5000000
10000000
15000000
20000000
25000000
30000000
mob2
Homozygous SNP density
Heterozygous SNP density
Hom/het ratio
0.00000000
0.00000004
0.00000008
0
5000000
10000000
15000000
20000000
25000000
30000000
sup#1
Homozygous SNP density
Heterozygous SNP density
Hom/het ratio
0.00000000
0.00000010
0.00000020
0.00000030
0
5000000
10000000
15000000
20000000

Figure 5

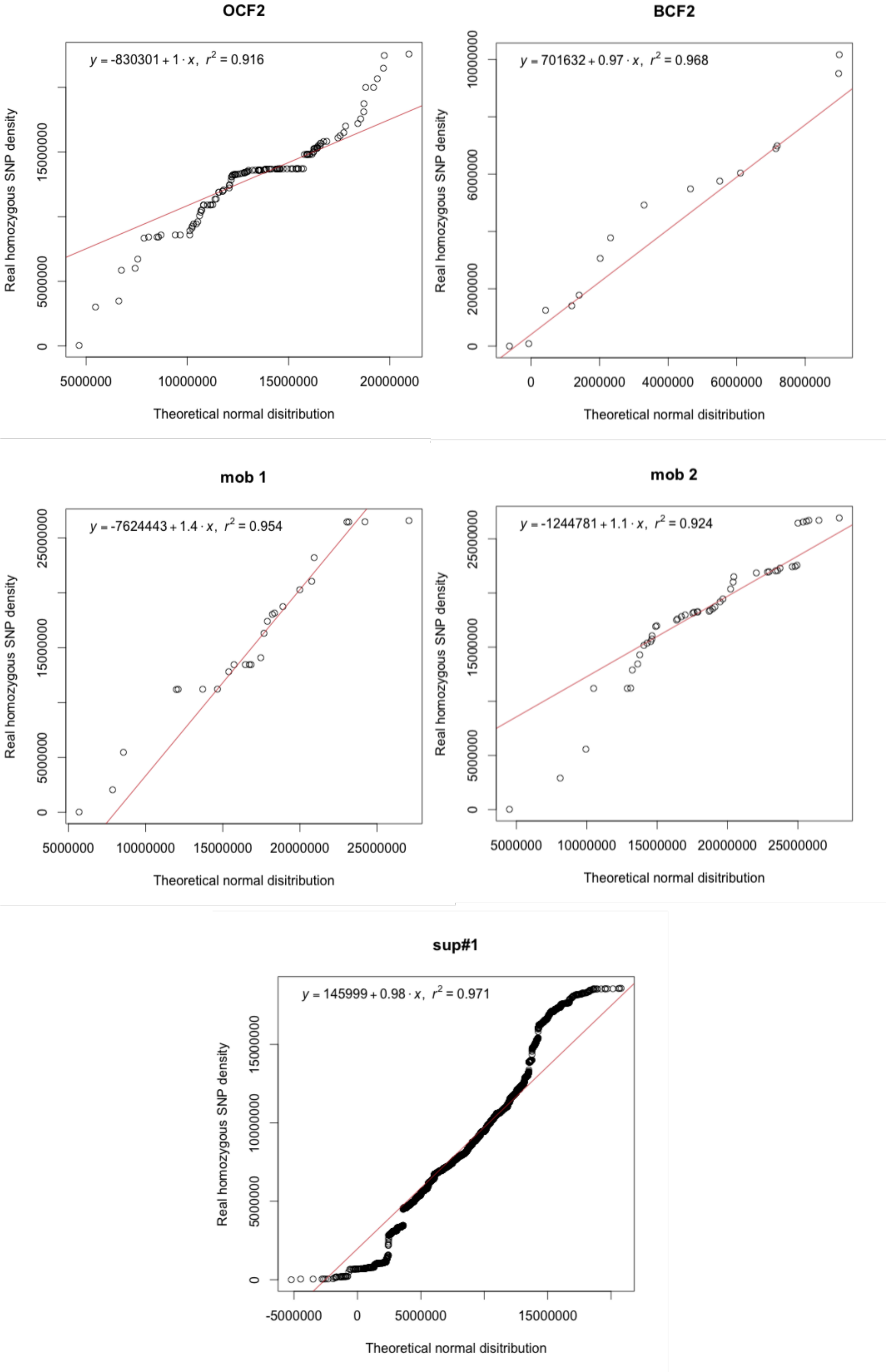
OCF2
$y = -830301 + 1 \cdot x,\ r^2 = 0.916$
Real homozygous SNP density
0
5000000
15000000
5000000
10000000
15000000
20000000
Theoretical normal disitribution
BCF2
$y = 701632 + 0.97 \cdot x,\ r^2 = 0.968$
Real homozygous SNP density
0
2000000
6000000
10000000
0
2000000
4000000
6000000
8000000
Theoretical normal disitribution
mob 1
$y = -7624443 + 1.4 \cdot x,\ r^2 = 0.954$
Real homozygous SNP density
0
5000000
15000000
25000000
5000000
10000000
15000000
20000000
25000000
Theoretical normal disitribution
mob 2
$y = -1244781 + 1.1 \cdot x,\ r^2 = 0.924$
Real homozygous SNP density
0
5000000
15000000
25000000
5000000
10000000
15000000
20000000
25000000
Theoretical normal disitribution
sup#1
$y = 145999 + 0.98 \cdot x,\ r^2 = 0.971$
Real homozygous SNP density
0
5000000
10000000
15000000
-5000000
0
5000000
15000000
Theoretical normal disitribution

Figure 6

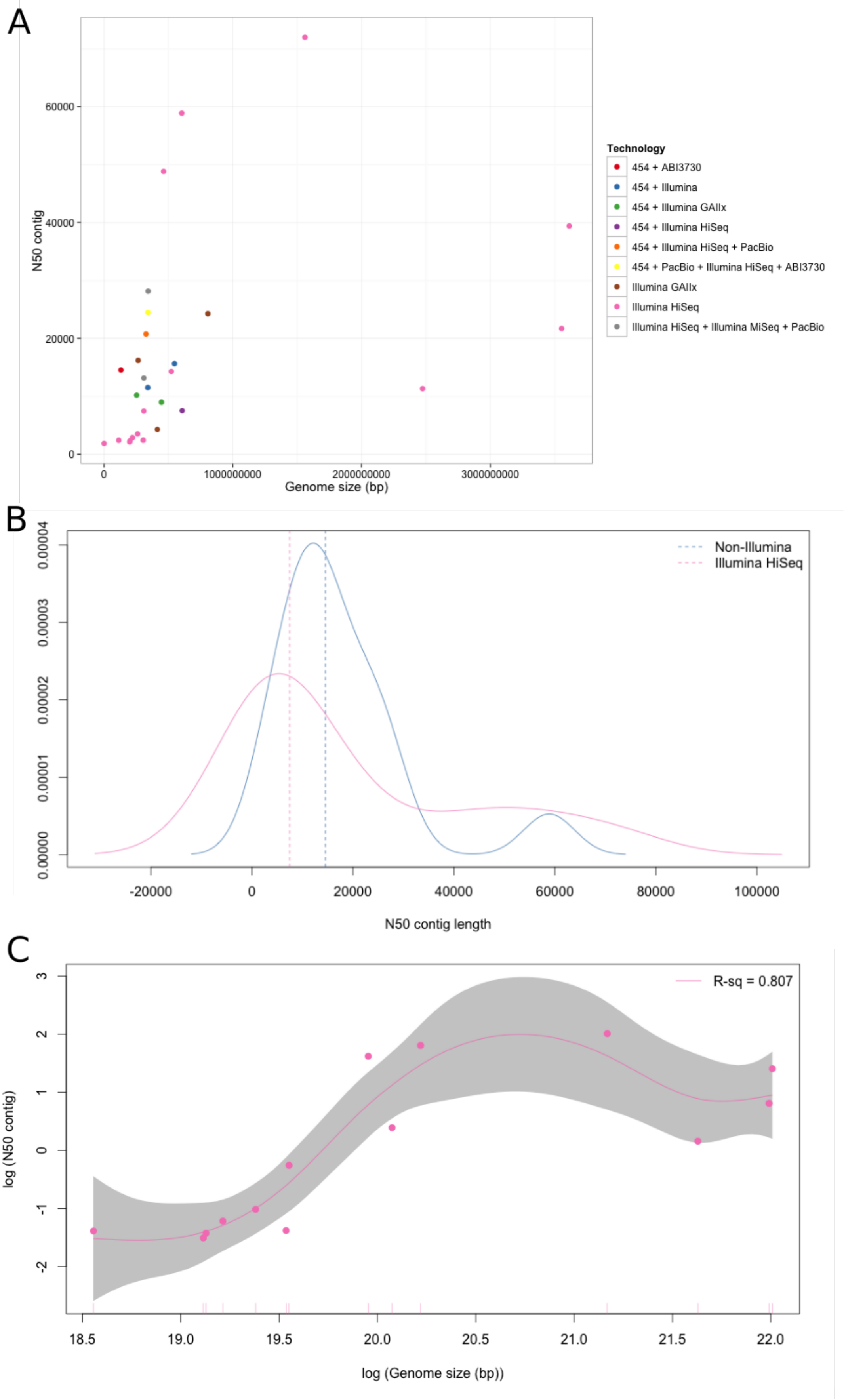
A
N50 contig
60000
40000
20000
0
0
1000000000
2000000000
3000000000
Genome size (bp)
Technology
● 454 + ABI3730
● 454 + Illumina
● 454 + Illumina GAIIx
● 454 + Illumina HiSeq
● 454 + Illumina HiSeq + PacBio
● 454 + PacBio + Illumina HiSeq + ABI3730
● Illumina GAIIx
● Illumina HiSeq
● Illumina HiSeq + Illumina MiSeq + PacBio
B
0.00004
0.00003
0.00002
0.00001
0.00000
Non-Illumina
Illumina HiSeq
-20000
0
20000
40000
60000
80000
100000
N50 contig length
C
log (N50 contig)
3
2
1
0
-1
-2
R-sq = 0.807
18.5
19.0
19.5
20.0
20.5
21.0
21.5
22.0
log (Genome size (bp))

Figure 7

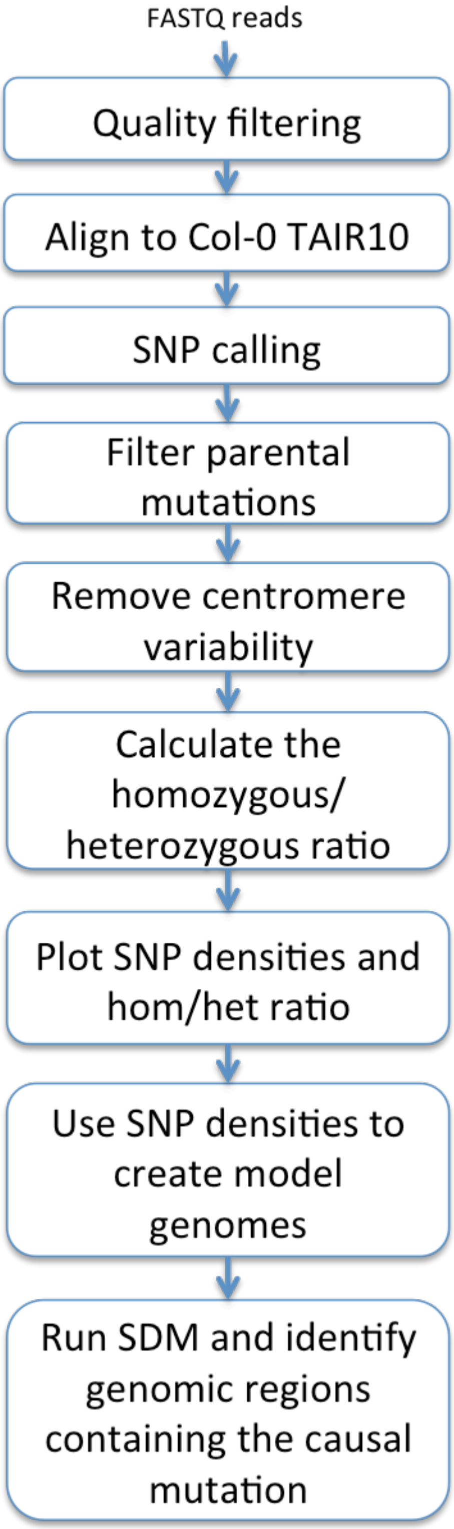
FASTQ reads
Quality filtering
Align to Col-0 TAIR10
SNP calling
Filter parental mutations
Remove centromere variability
Calculate the homozygous/ heterozygous ratio
Plot SNP densities and hom/het ratio
Use SNP densities to create model genomes
Run SDM and identify genomic regions containing the causal mutation

Figure 8

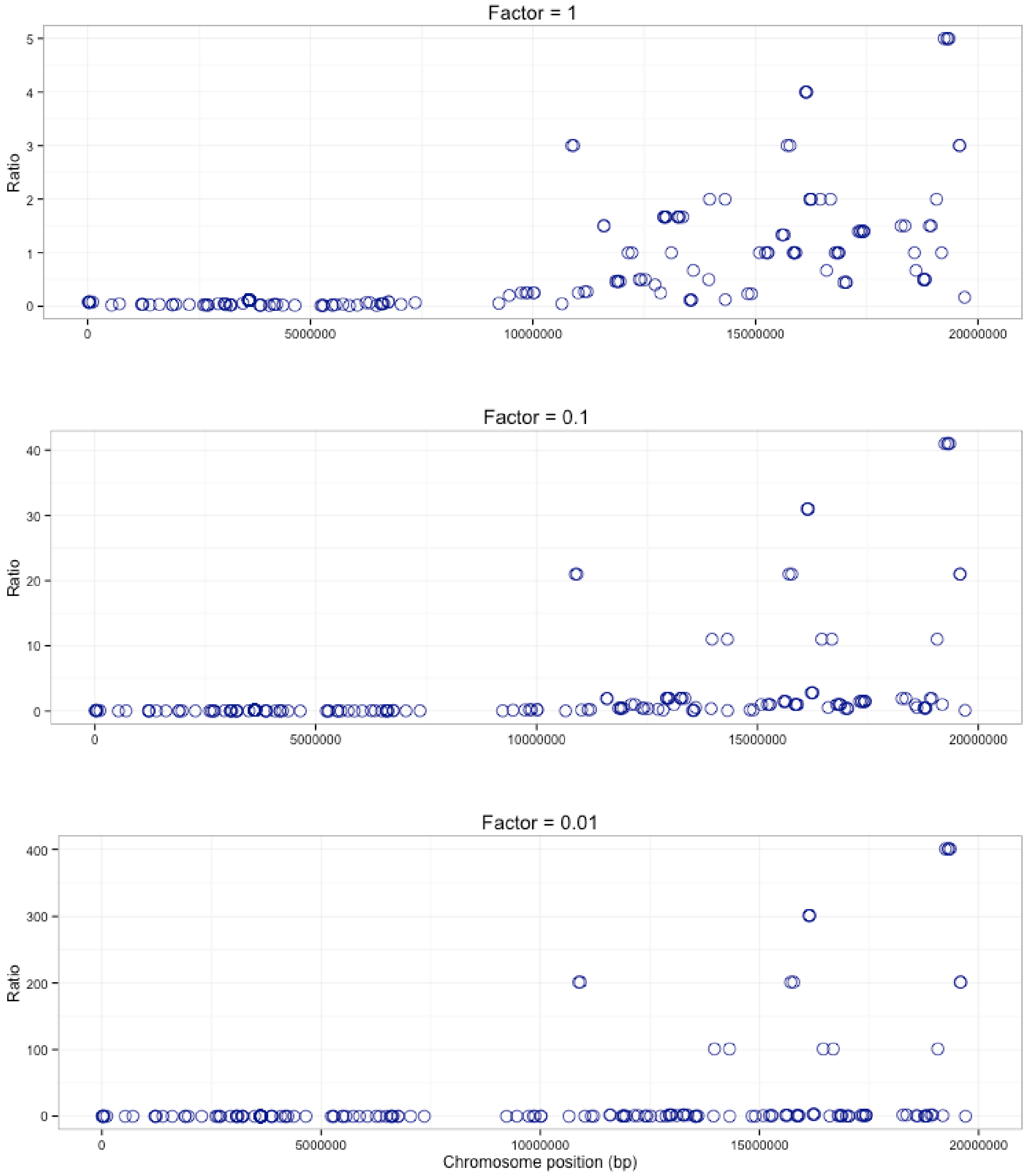
Factor = 1
Ratio
5
4
3
2
1
0
0
5000000
10000000
15000000
20000000
Factor = 0.1
Ratio
40
30
20
10
0
0
5000000
10000000
15000000
20000000
Factor = 0.01
Ratio
400
300
200
100
0
0
5000000
10000000
15000000
20000000
Chromosome position (bp)

Figure 9

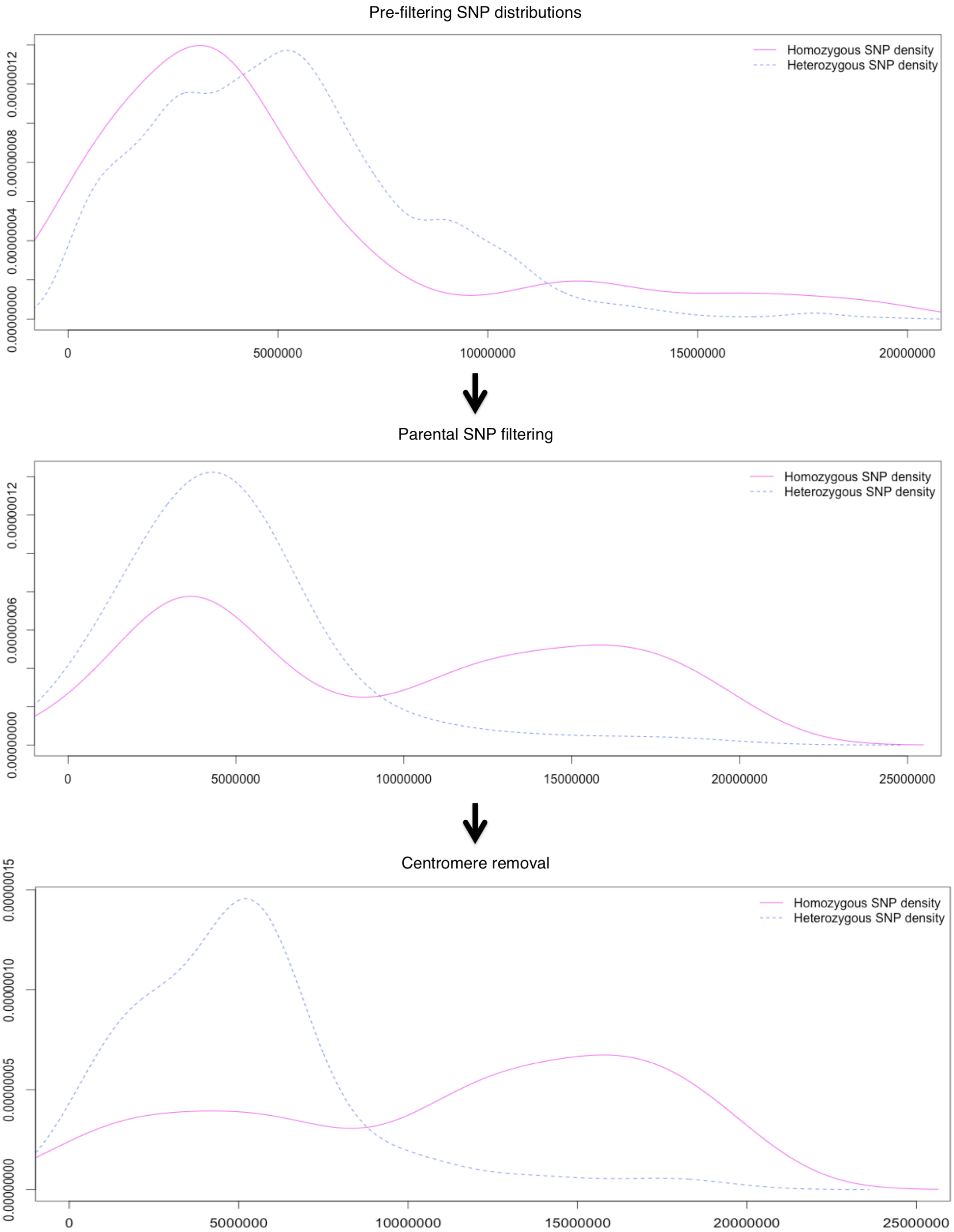
Pre-filtering SNP distributions
Homozygous SNP density
Heterozygous SNP density
0.00000000
0.00000004
0.00000008
0.00000012
0
5000000
10000000
15000000
20000000
Parental SNP filtering
Homozygous SNP density
Heterozygous SNP density
0.00000000
0.00000006
0.00000012
0
5000000
10000000
15000000
20000000
25000000
Centromere removal
Homozygous SNP density
Heterozygous SNP density
0.00000000
0.00000005
0.00000010
0.00000015
0
5000000
10000000
15000000
20000000
25000000

Figure 10

```
   sh 'samtools view -bS alignment.sam | samtools sort -m 30000000000 - alignment'
end
```

4. Generate pileup from BAM file

```
desc "Write pileup file"
task :pileup => ["bam"] do
     sh 'samtools mpileup -B -f TAIR10.fa alignment.bam > SNPs.pileup'
end
```

5. Call SNPs using VarScan and record them in a VCF4.1 file

```
desc "run VarScan"
task :varscan  => ["pileup"] do
     sh 'java -jar VarScan.v2.3.7.jar mpileup2snp SNPs.pileup --output-vcf 1
        > SNPs.vcf'
end
```

sup_file1.md Supplementary file 1. Method workflow ===

Quality filtering for paired-end reads

```
FastQC/fastqc reads_R1.fq
FastQC/fastqc reads_R2.fq
```

```
java -jar Trimmomatic-0.33/trimmomatic-0.33.jar PE reads_R1.fq reads_R2.fq paired_R1
```

Command line parameters used for paired-end read mapping and SNP calling

1. Index reference sequence

```
bwa index TAIR10.fa
```

2. Map the reads to reference genome with BWA

```
desc "Align using bwa"
task :bwa  do
     sh 'bwa mem TAIR10.fa paired_R1.fq paired_R2.fq > alignment.sam'
end
```

3. Convert the respective SAM file to BAM file and sort the BAM file using samtools

```
desc "Convert sam to bam file"
task :bam => ["bwa"] do
   sh 'samtools view -bS alignment.sam | samtools sort -m 30000000000 - alignment'
end
```

4. Generate pileup from BAM file

```
desc "Write pileup file"
task :pileup => ["bam"] do
     sh 'samtools mpileup -B -f TAIR10.fa alignment.bam > SNPs.pileup'
end
```

5. Call SNPs using VarScan and record them in a VCF4.1 file

```
desc "run VarScan"
task :varscan  => ["pileup"] do
     sh 'java -jar VarScan.v2.3.7.jar mpileup2snp SNPs.pileup --output-vcf 1
        > SNPs.vcf'
end
```