Figures

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. Effect 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)** 1% of the maximum ratio. **(B)** 5% of the maximum ratio. **(C)** 10% of the maximum ratio. **(D)** 20% 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. Overlapping homozygous and heterozygous SNP densities and hom/het 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 verify the correlation.

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 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
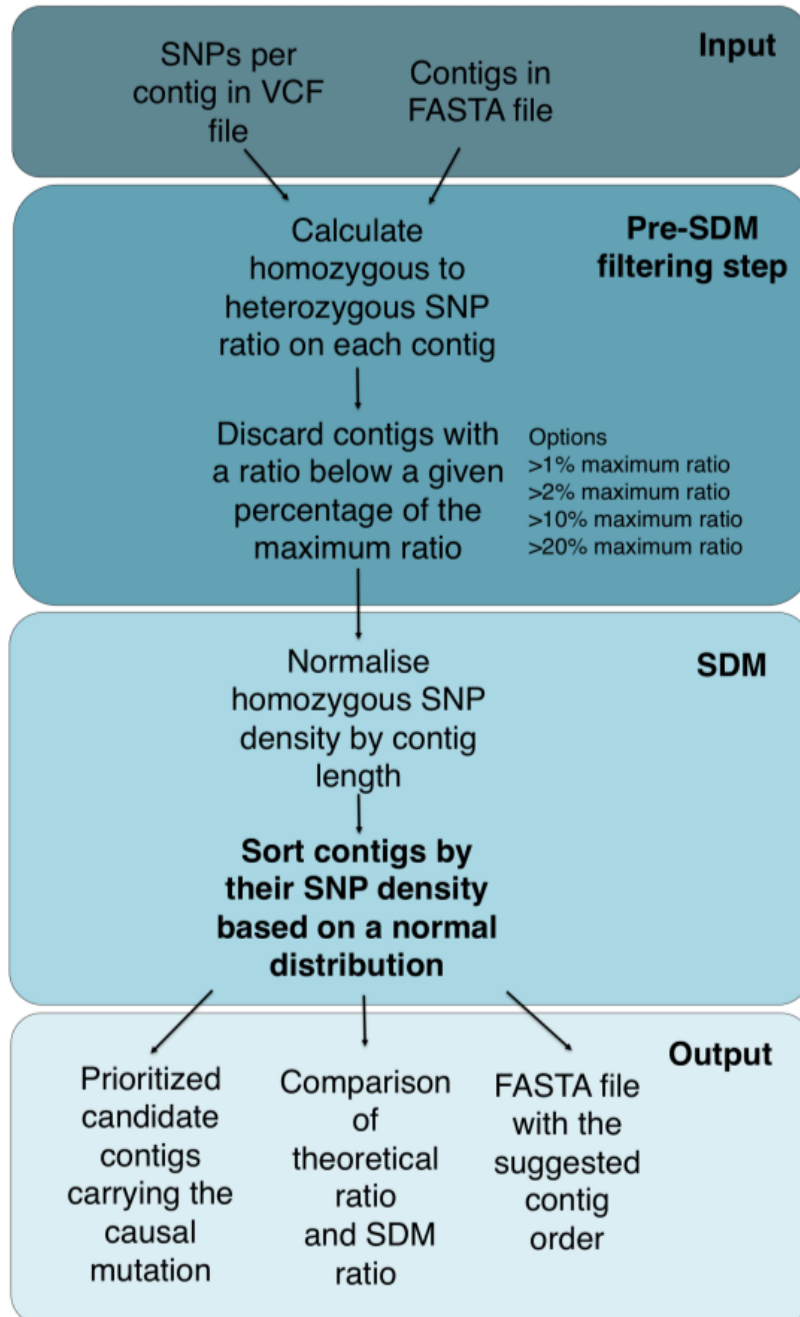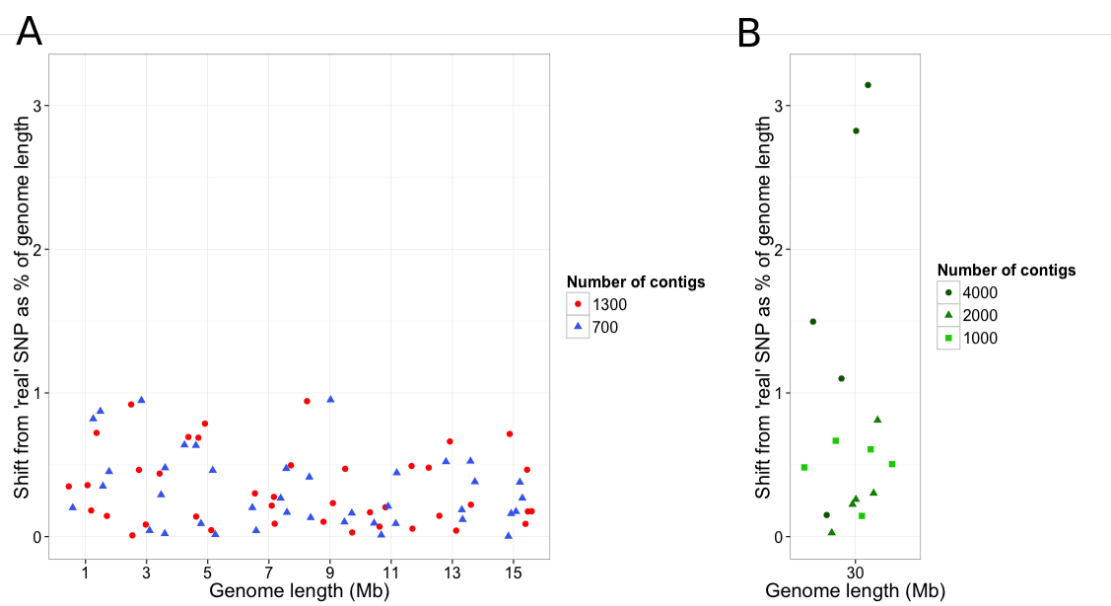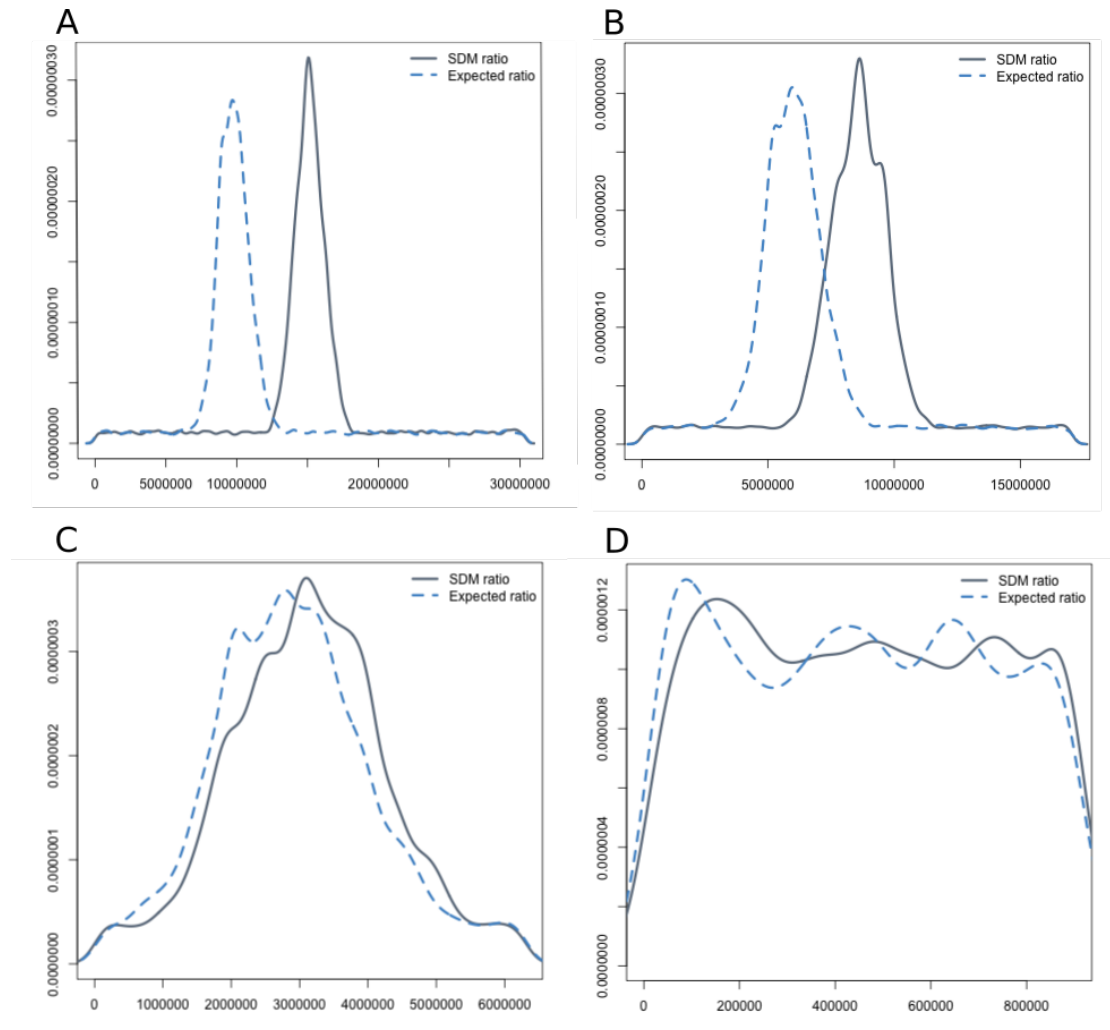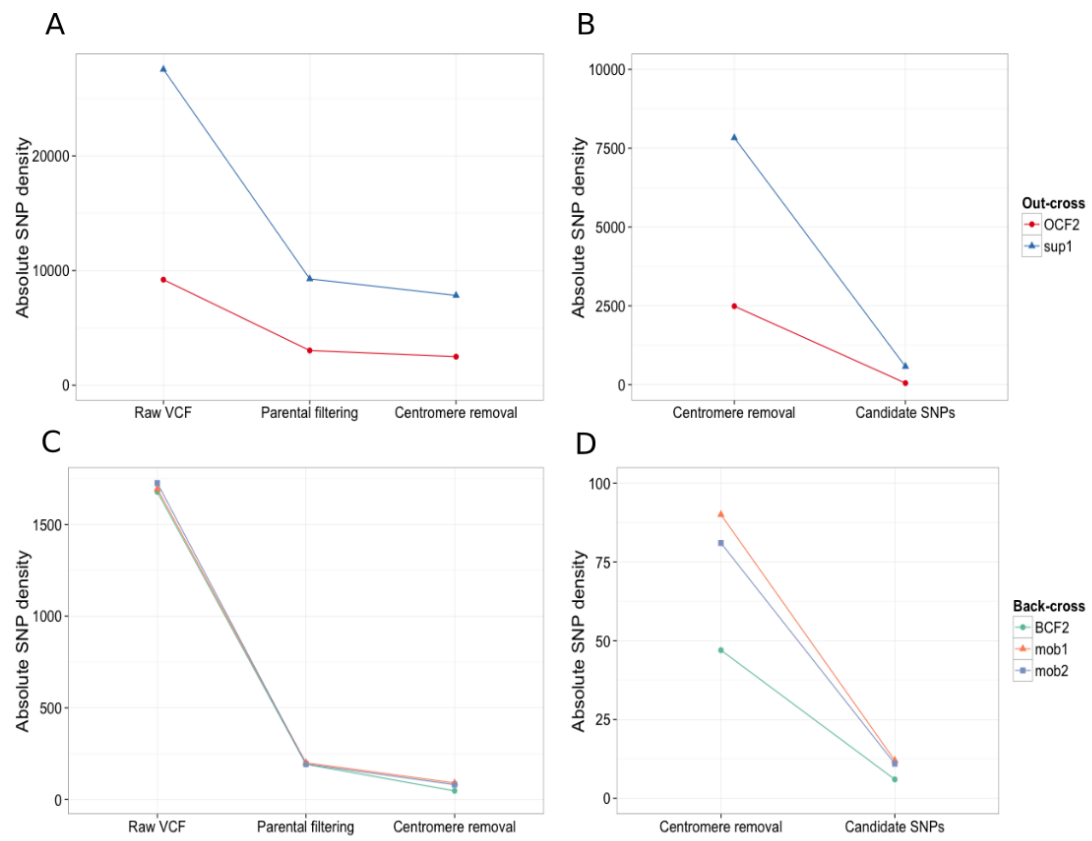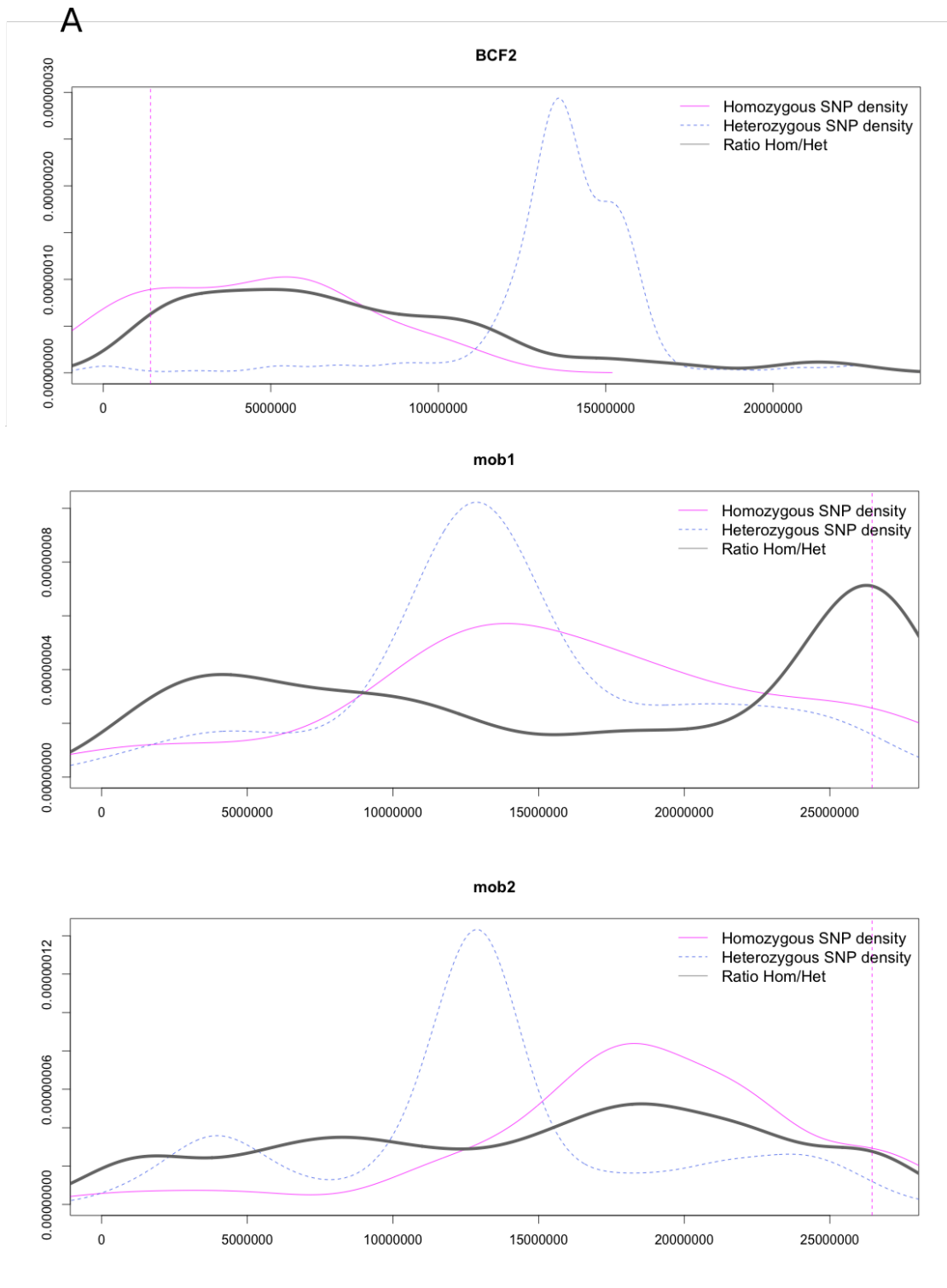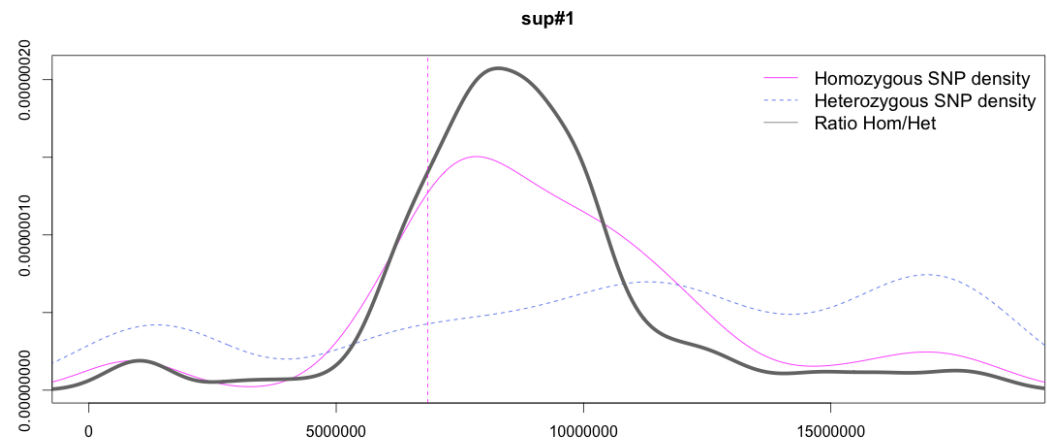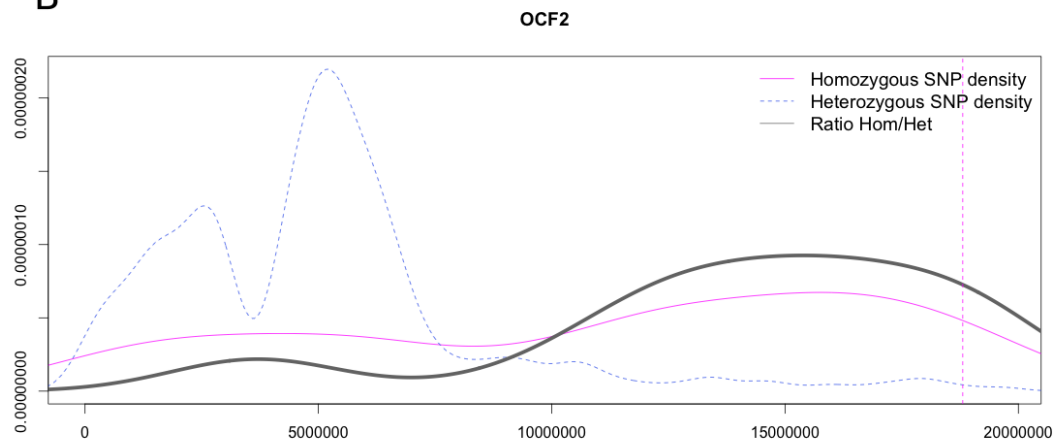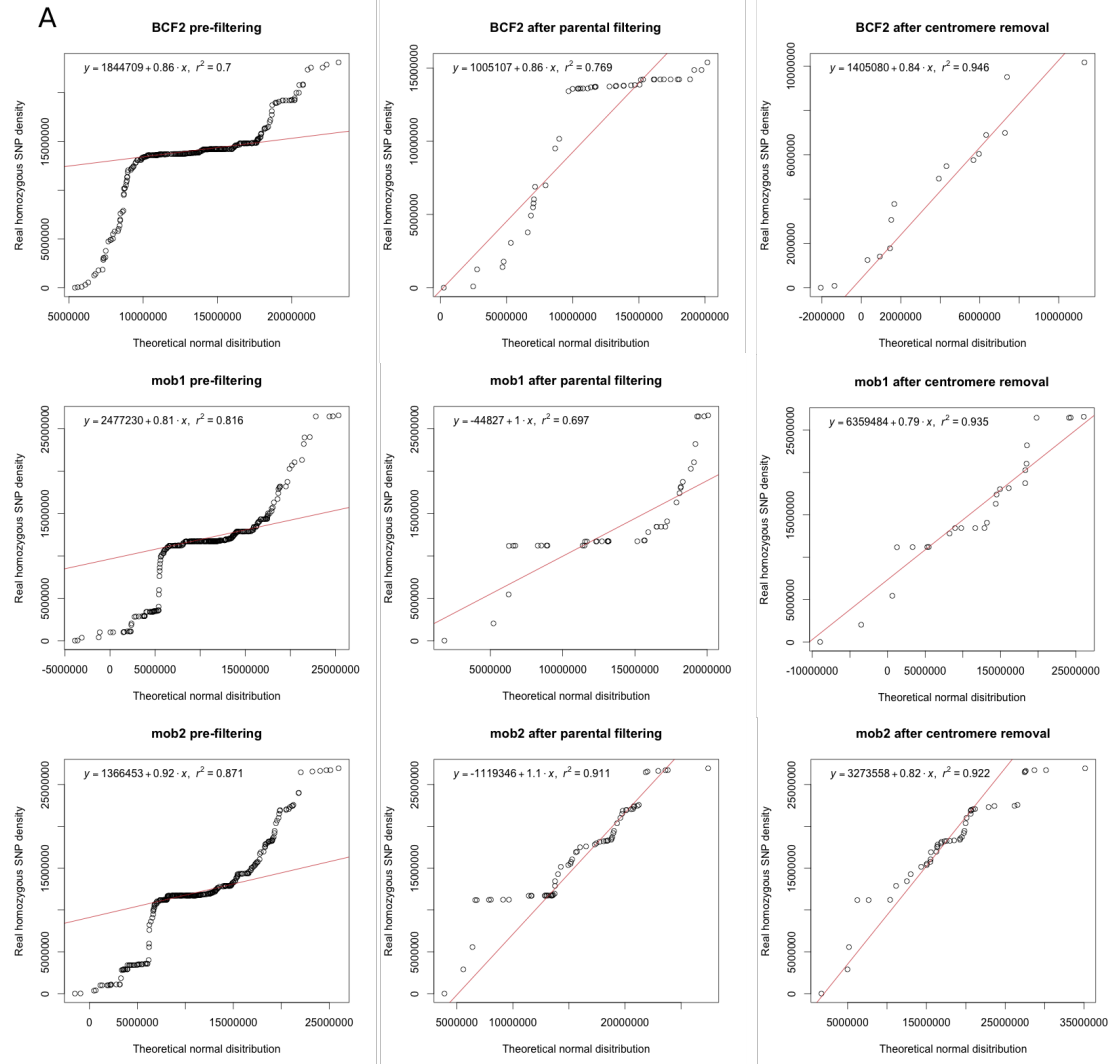
Figure 1

Figure 2

Figure 3

Figure 4

A

B

C

D

Figure 5

A

BCF2

mob1

mob2

Figure 5

Figure 6

A

BCF2 pre-filtering

$y = 1844709 + 0.86 \cdot x, \ r^2 = 0.7$

BCF2 after parental filtering

$y = 1005107 + 0.86 \cdot x, \ r^2 = 0.769$

BCF2 after centromere removal

$y = 1405080 + 0.84 \cdot x, \ r^2 = 0.946$

mob1 pre-filtering

$y = 2477230 + 0.81 \cdot x, \ r^2 = 0.816$

mob1 after parental filtering

$y = -44827 + 1 \cdot x, \ r^2 = 0.697$

mob1 after centromere removal

$y = 6359484 + 0.79 \cdot x, \ r^2 = 0.935$

mob2 pre-filtering

$y = 1366453 + 0.92 \cdot x, \ r^2 = 0.871$

mob2 after parental filtering

$y = -1119346 + 1.1 \cdot x, \ r^2 = 0.911$

mob2 after centromere removal

$y = 3273558 + 0.82 \cdot x, \ r^2 = 0.922$

Real homozygous SNP density

Theoretical normal disitribution

Figure 6

B

OCF2 pre-filtering

$y = 682263 + 0.86 \cdot x,\ r^2 = 0.806$

OCF2 after parental filtering

$y = 464720 + 0.9 \cdot x,\ r^2 = 0.889$

OCF2 after centromere removal

$y = -820720 + 1.1 \cdot x,\ r^2 = 0.949$

sup1 pre-filtering

$y = 186720 + 0.98 \cdot x,\ r^2 = 0.977$

sup1 after parental filtering

$y = 67164 + 0.99 \cdot x,\ r^2 = 0.968$

sup1 after centromere removal

$y = 19345 + 1 \cdot x,\ r^2 = 0.978$

Figure 7

A

Figure 7