New sequencing data: Early results

Cyril Matthey-Doret

08/10/2017

General statistics

Number of individuals

From the 768 processed samples, 704 made it to the end of the pipeline. Out of those, there were 365 haploid males. Among the 339 diploid individuals left, there are 176 males and 163 females. There is a total of 60 families among which 48 contain diploid individuals and 37 contain both diploid males and females.

Homozygosity values

Proportion of homozygosity among variant sites was used to classify individuals as haploid or diploid. The figure below represents the distribution of homozygosity. across all individuals

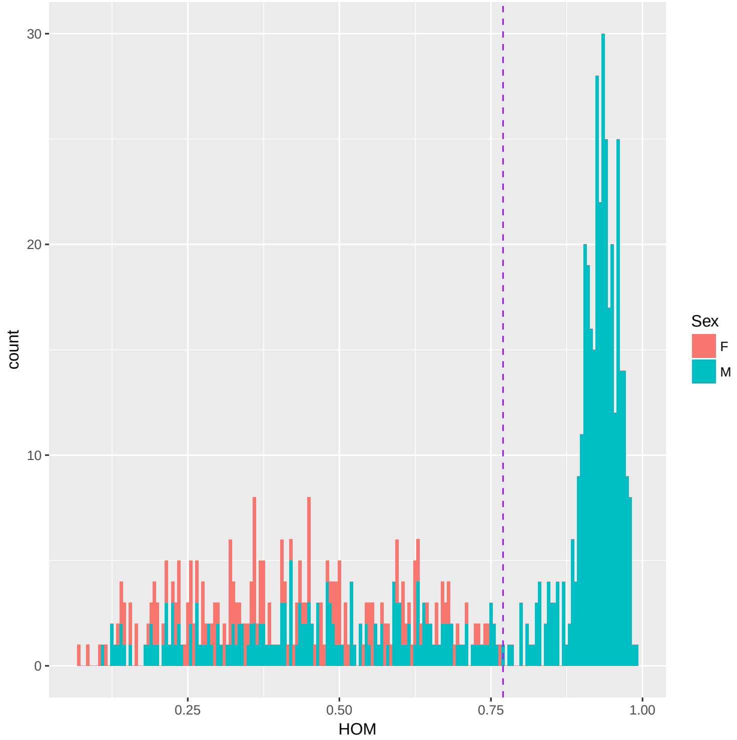

Figure 1: Distribution of the proportion of homozygosity among variant sites per individuals.

Grouping families

The goal in this section is to split families into groups that reflect the CSD genotype of the mothers. I attempted to achieve this by clusterint families according to the proportion of males among diploid offspring.

Because we only sequence a fraction of the offspring in each family, I computed the rate or haploidy among sequenced males in each family and extrapolated this proportion to all (non-sequenced) males to infer the total number of diploid males. I then used these values to measure the ratio of diploid males to diploid offspring.

The expected number of groups depend on the number of CSD loci we consider. Each group will reflect the number of heterozygous CSD loci in the mothers as follows.

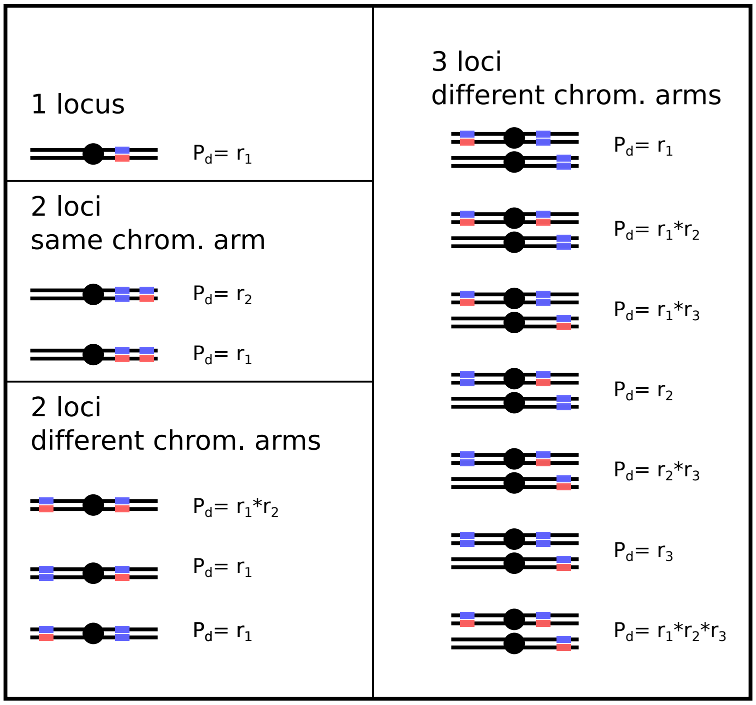

Figure 2: Possible mother genotypes of mothers with associated proportion of diploid males when considering different number of CSD loci. The proportion of diploid males produced by a mother (Pd) depends on the recombination rate r_x at the heterozygous CSD loci.

I clustered families automatically using the k-means methods in order to obtain unbiased groups. I considered 2 different scenarios: 2 CSD loci on different chromosomal arms and 3 CSD loci on different chromosomal arms. I don't have enough families to detect scenarios with more loci. The clusters are not clearly separated and this could be due to: inaccurate metric, high number of loci, high variability in recombination rates… ?

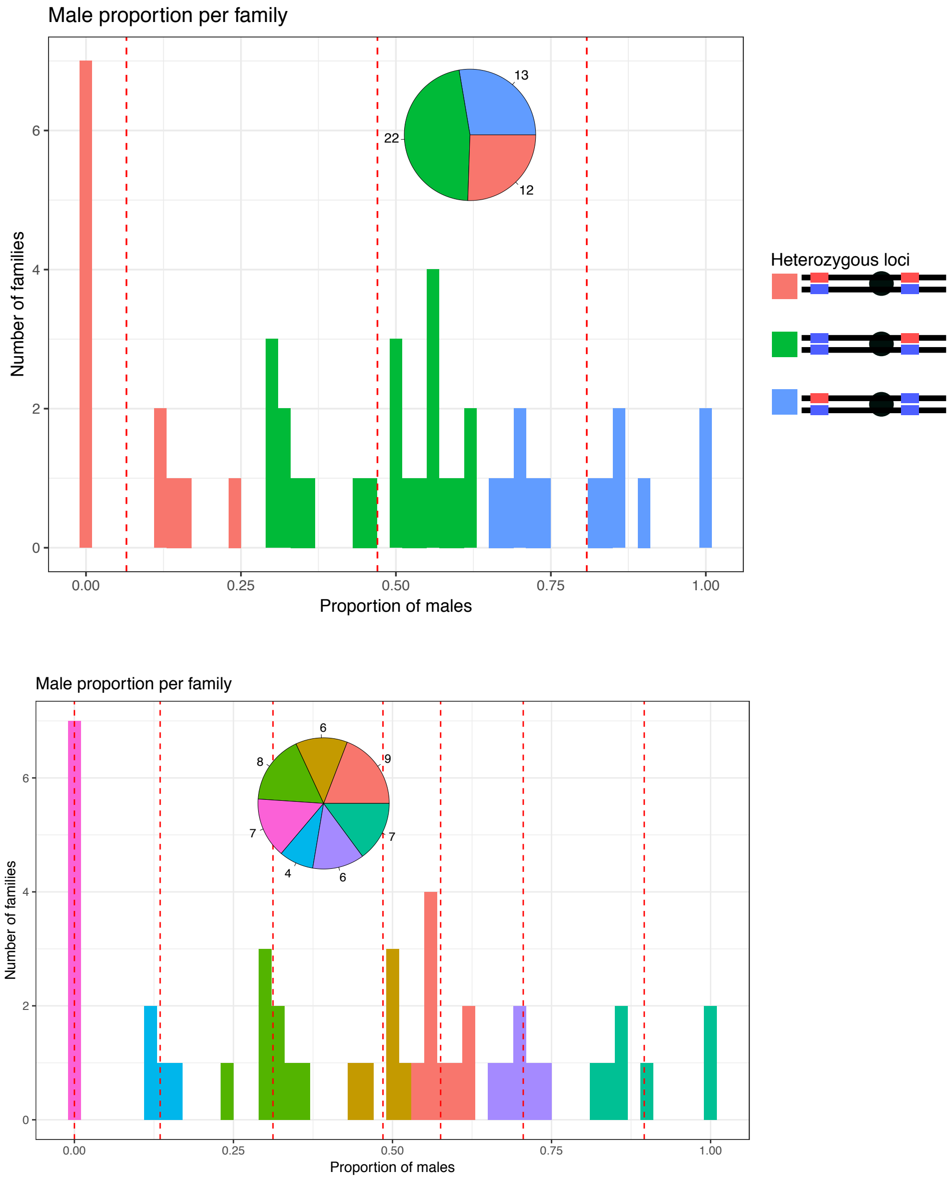

Figure 3: Family groups obtained through k-means clustering based on inferred proportion of males among diploid offspring. top: scenario considering 2 loci on different chromosomal arms, genotypes corresponding to the clusters are displayed in the legend. bottom: scenario considering 3 loci on different chromosomal arms. Genorypes are not displayed as they will depend on the relative recombination rates of the loci. Piecharts show the number of families belonging to each cluster.

Association mapping

I performed association mapping both pooling all families together, or splitting them according to the two scenario above. I use one-sided Fisher exact test on each locus to test if there is a there is a positive association between heterozygosity at the locus and female sex.

2 loci

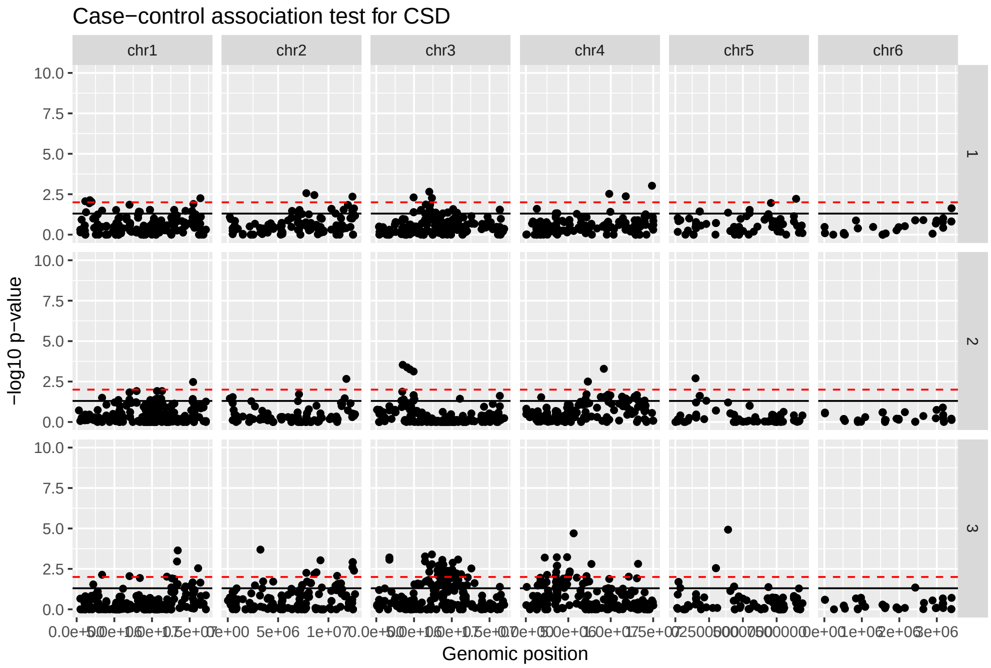

Figure 4: Association mapping considering 2 loci scenario. No correction for multiple testing was applied.

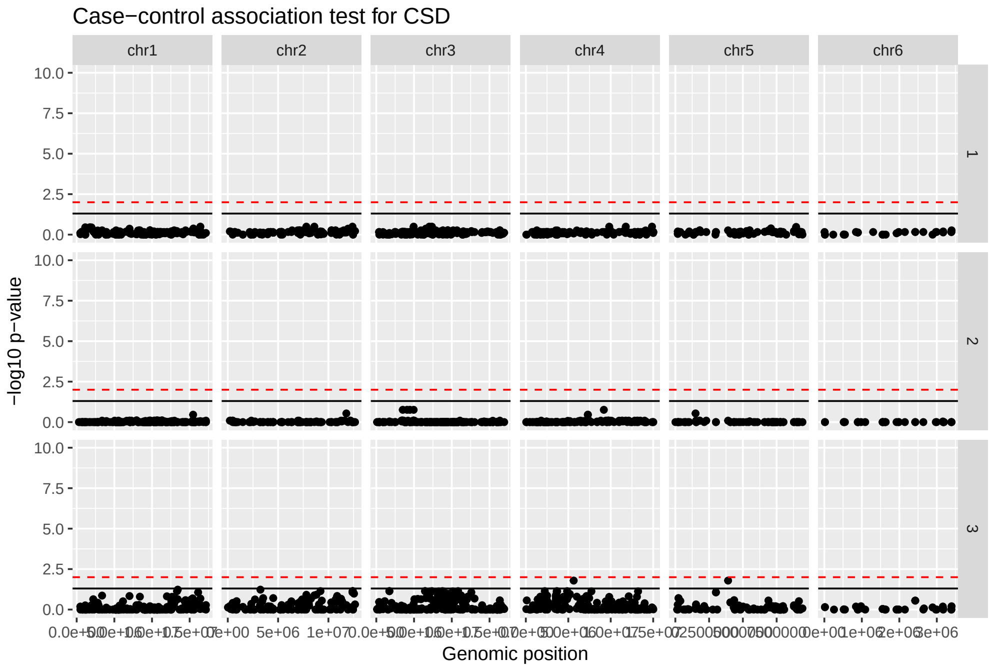

Figure 5: Association mapping considering 2 loci scenario. Benjamini-Hochberg correction for multiple testing was applied.

3 loci

Note that group 7 is always absent because it only contains haploid mals and daughters. -> sexual mothers ?

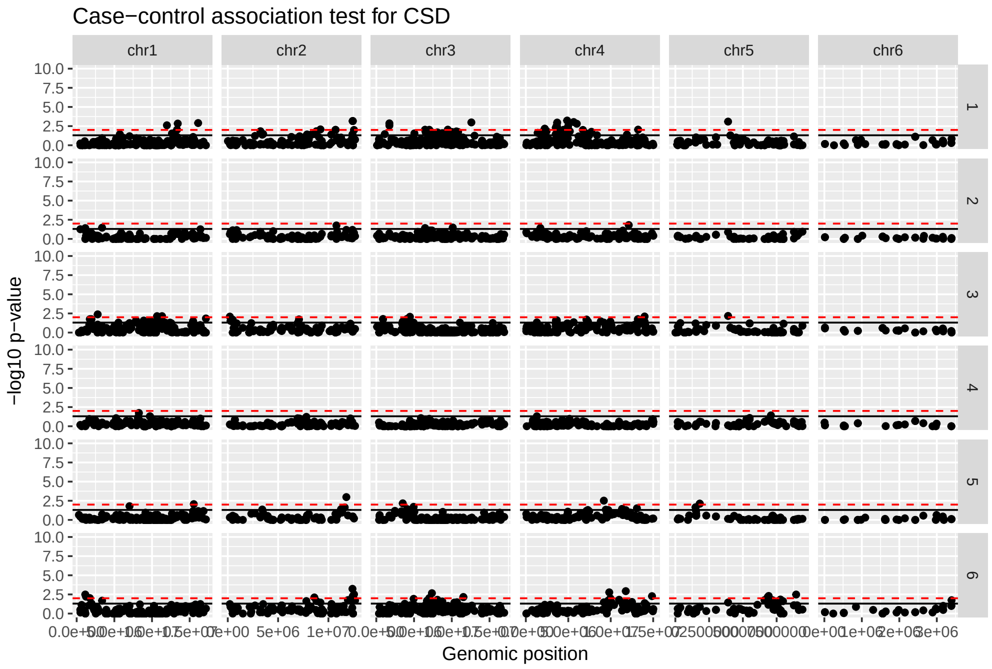

Figure 6: Association mapping considering 3 loci scenario. No correction for multiple testing was applied.

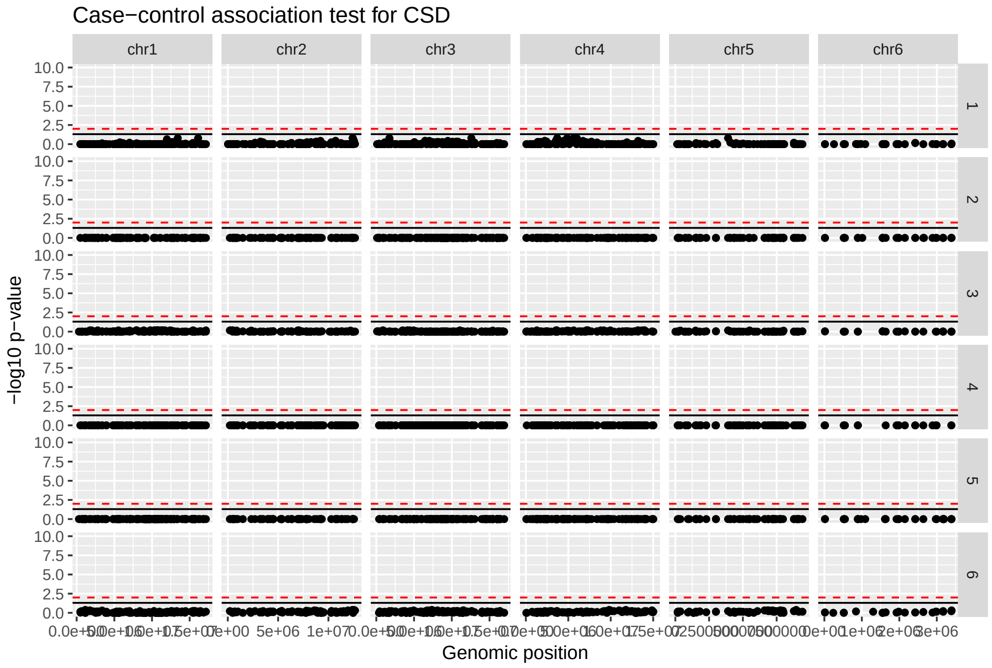

Figure 7: Association mapping considering 3 loci scenario. Benjamini-Hochberg correction for multiple testing was applied.

Pooling all families

Non-mapped contigs

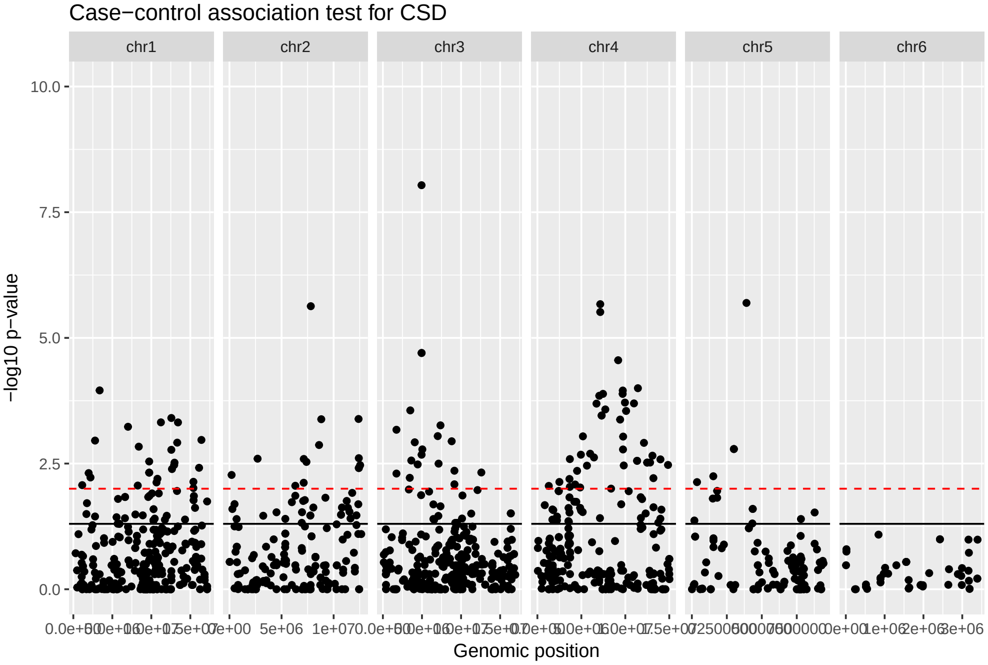

Figure 8: Association mapping pooling all families together. No correction for multiple testing was applied.

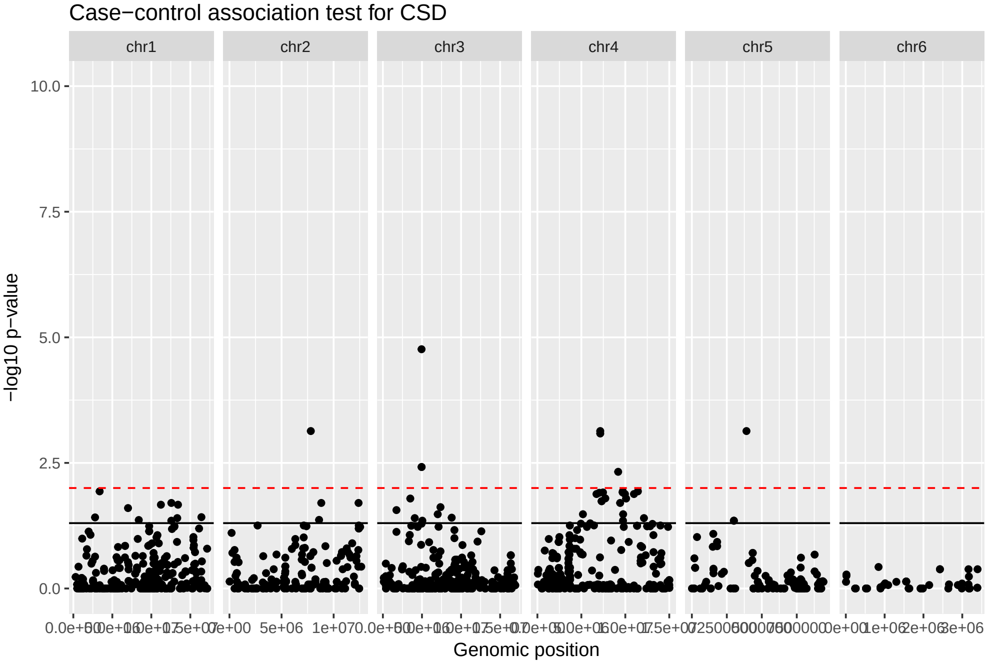

Figure 9: Association mapping pooling all families together. Benjamini-Hochberg for multiple testing was applied.

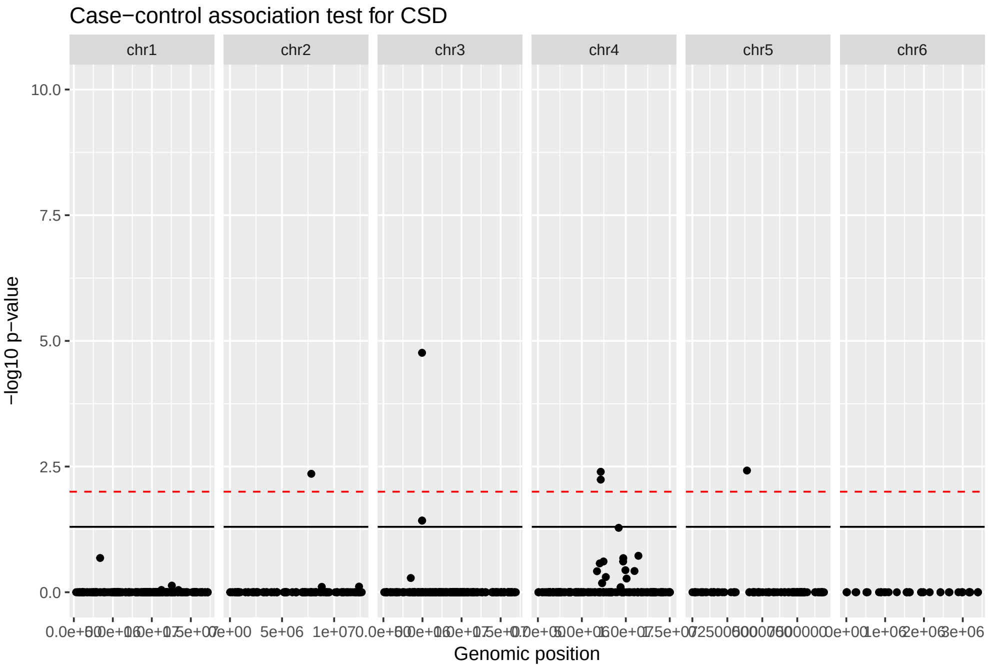

Figure 10: Association mapping pooling all families together. Bonferroni for multiple testing was applied.

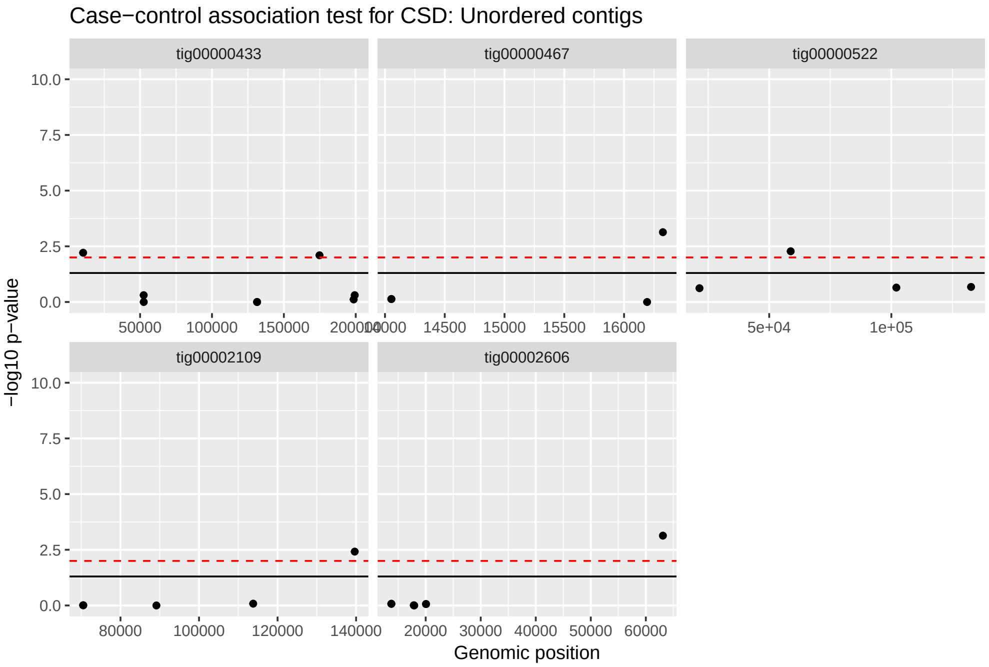

Figure 11: Association mapping pooling all families together showing unordered contigs only. Benjamini-Hochberg correction for multiple testing was applied.

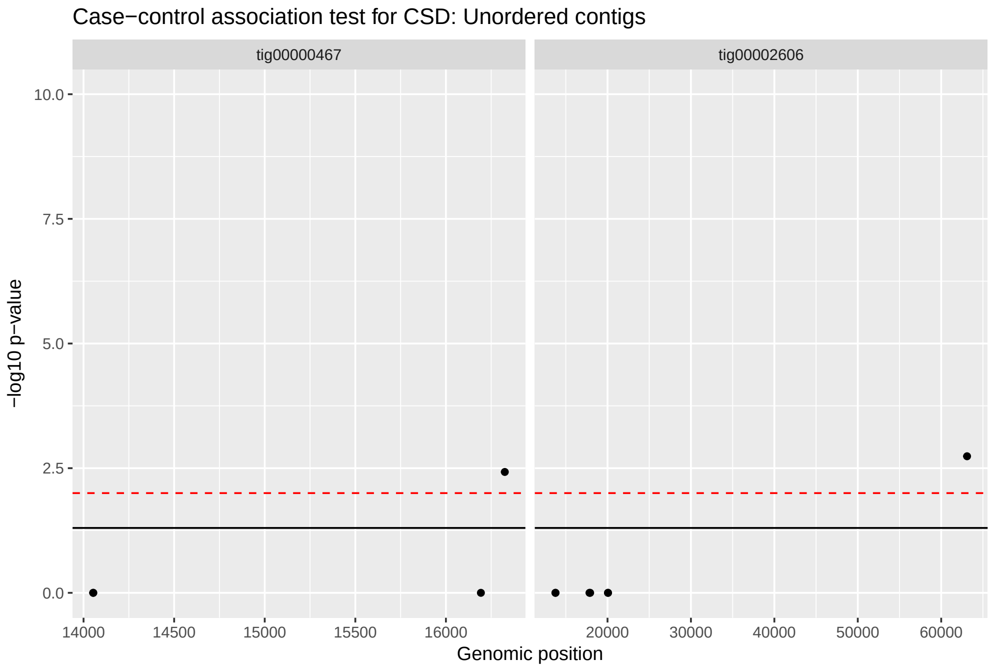

Figure 12: Association mapping pooling all families together showing unordered contigs only. Bonferroni for multiple testing was applied.

Ideas

More candidates than expected when pooling families, none when clustering. The clustering is probably wrong and different genotypes are mixed in the same groups.

What I could do: Blast candidate regions. Look at what kind of genes are in there. Check for homologies across regions and with other hymenoptera.

Possible to improve genome using my samples ? Linkage map etc (ALLMAPS)

Fix clustering problem using different metric ? (but what, then ?)

Pipeline parameters

mapping: BWA aln, Mismatches=4

tot	single	multi	miss
170435458	0.555	0.261	0.183

pstacks: mininum stack coverage=3

Stats averaged across samples.

N loci	mean coverage	SD coverage
2694	50	69.4

cstacks: Mismatch=3

N Loci	N alleles
12585	56971

populations: Min proportion (R)=0.8, min depth=5

mean depth	median depth	sd depth	mean N sites	median N sites	sd N sites
80.56479	84.4192	34.6024	2260.802	2296	97.14321