Last Sequencing data: Summary

Cyril Matthey-Doret

16/10/2017

Context

After receiving the second run of sequencing data for libraries 12 and 12b, I pooled them for demultiplexing and ran the usual STACKS pipeline per-family with stringent parameter for splitting by ploidy (min locus depth 20). The one thing I changed in the pipeline is that I removed the max_obs_het filter which excluded loci heterozygous in more than 90% of individuals. I will replace this filter by a blacklist of loci heterozygous in more than *50%* of haploid males after inference of ploidy. The distribution of homozygosity per individuals revealed a highly homozygous cluster of individuals from the D cross forming a second mode close to the haploids (Figure 1). This is problematic because it will be hard to disentangle it from the haploid mode. After discussing with Casper, we concluded that it would be more appropriate to consider the D-cross separately as it could also cause the catalog to exclude important loci in the C cross (due to the R=0.8 filter) from the catalog.

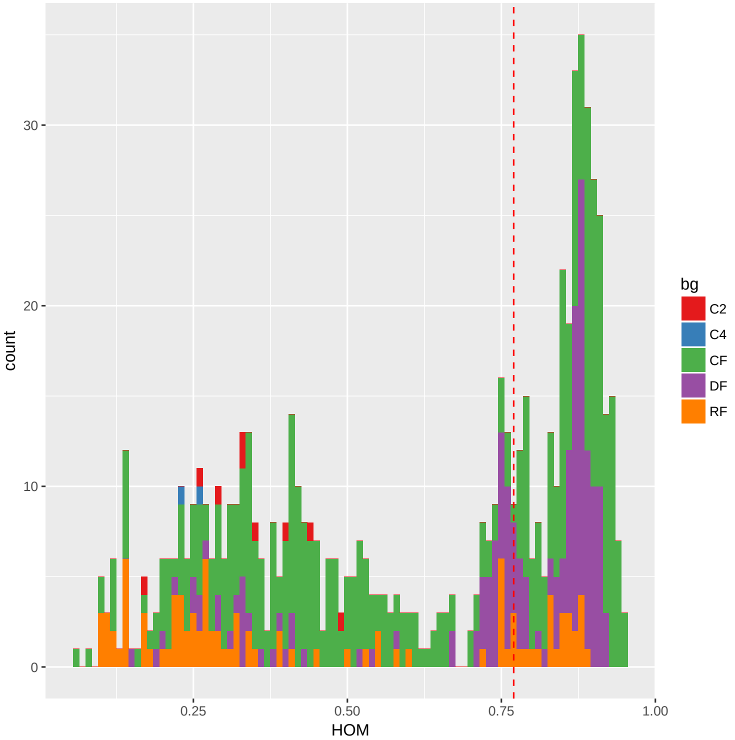

Figure 1: Distribution of the proportion of homozygosity among variant sites per individuals. Populations was run per-family without max-obs-het filter. Individuals are colored based on background.

First results

Before running the D cross in a separate catalog, I tried running the populations program with all individuals together instead of per-family. It did remove most of the second mode (Figure 2) which now appears to be merged with the haploid one. This was just to assess the impact of per-family versus pooled populations run on homozygosity, but we will still exclude the D cross from the main pipeline, however. As it appears, most individuals in the high-homozygosity mode were in fact haploids, and removing this cross from the analysis will have little to no effect as it contains very few females.

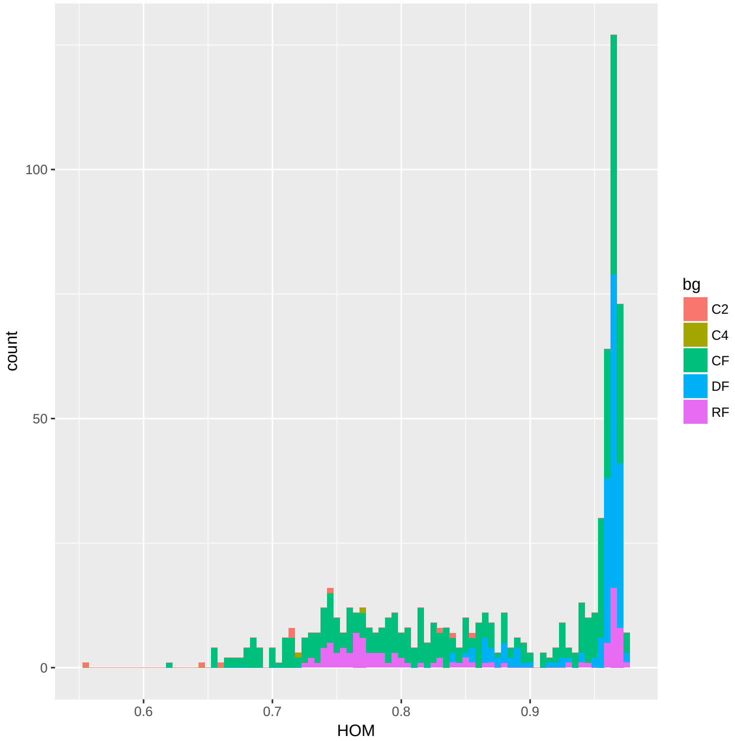

Figure 2: Distribution of the proportion of homozygosity among variant sites per individuals. Populations was run on all samples together without max-obs-het filter. Individuals are colored based on background.

Splitting crosses

C cross

This cross contains 588 individuals, of which 530 made it through the pipeline. They are split into 45 families.

- Mean number of sites per sample: 1439
- Average locus depth per sample: 127.
- Ploidy counts: 314 diploids, 216 haploids

D cross

This cross contains 186 individuals, of which 174 made it through the pipeline. They are split into 15 families.

- Mean number of sites per sample: 864
- Average locus depth per sample: 108
- Ploidy count: 35 diploids, 139 haploids

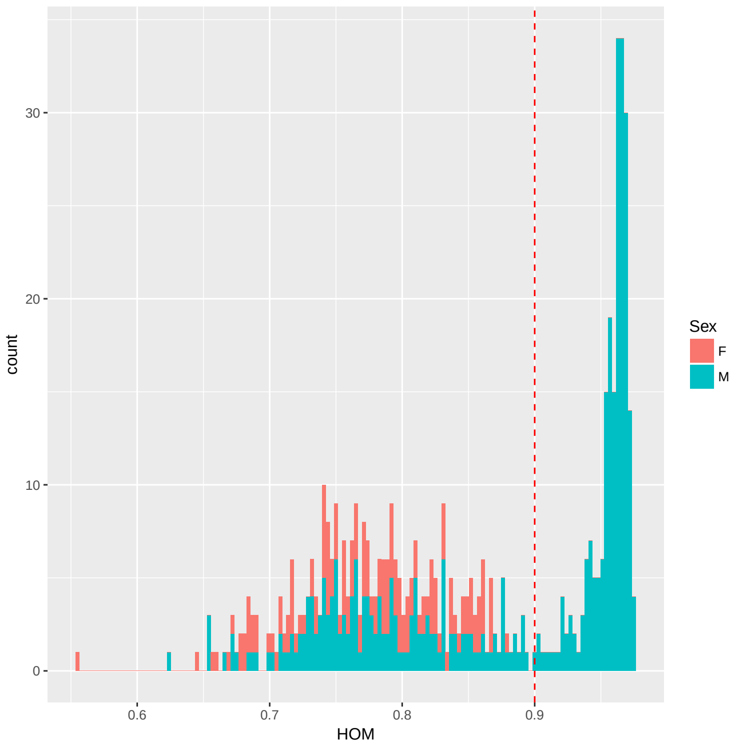

Figure 3: Distribution of the proportion of homozygosity among variant sites per individuals. Populations was run on all samples from cross C together without max-obs-het filter. Individuals are colored based on background.

Association mapping

The families still don't cluster by proportion of males among diploids, but when pooling them all together, it seems there are 3 good candidate regions (Figures 5, 6, 7).

Method changes

Up to now, I ran populations per family to split individuals by ploidy to improve the number of loci per individuals, however this practice seems to generate highly different levels of homozygosity between families and it would be less risky to pool individuals together and have a larger overlapping set of loci between families if I want to use a universal cutoff to split by ploidy.

I am now excluding loci that are found to be heterozygous in haploid males and this seems like a better strategy than using maximum proportion of heterozygous individuals. The proportion of haploids which need to be heterozygous at the locus might be subject to change however.

Brainstorming

Wavelet transform to estimate size of centromeric region

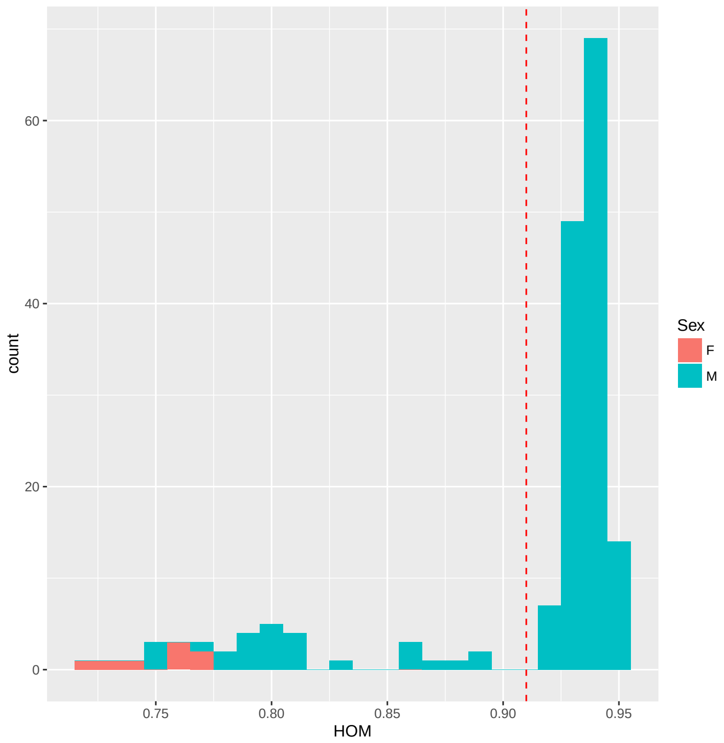

Figure 4: Distribution of the proportion of homozygosity among variant sites per individuals. Populations was run on all samples from cross D together without max-obs-het filter. Individuals are colored based on background.

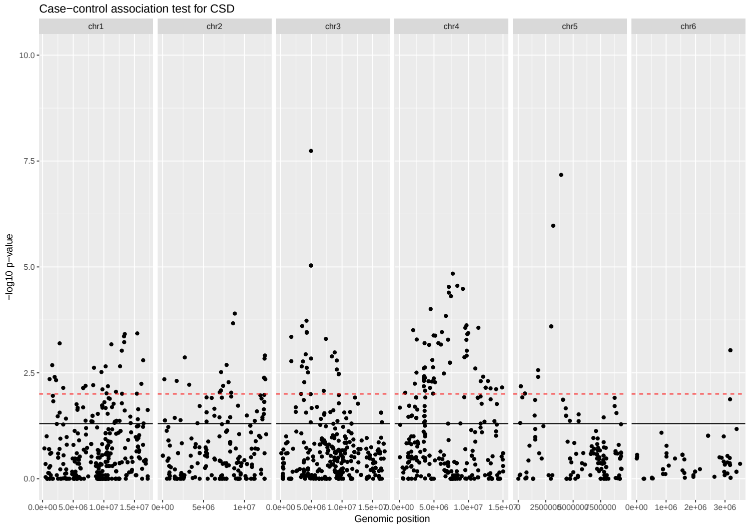

Figure 5: Manhattan plot of the case-control association mapping with fisher exact test. No correction for multiple testing was applied. Red dashed line: p=0.01, black continuous line: p=0.05

Planning: Priority

1. Annotations in candidate regions -> poster / presentation
2. Linkage map: Impossible to make a new one -> use Jens' to overlay centromeres

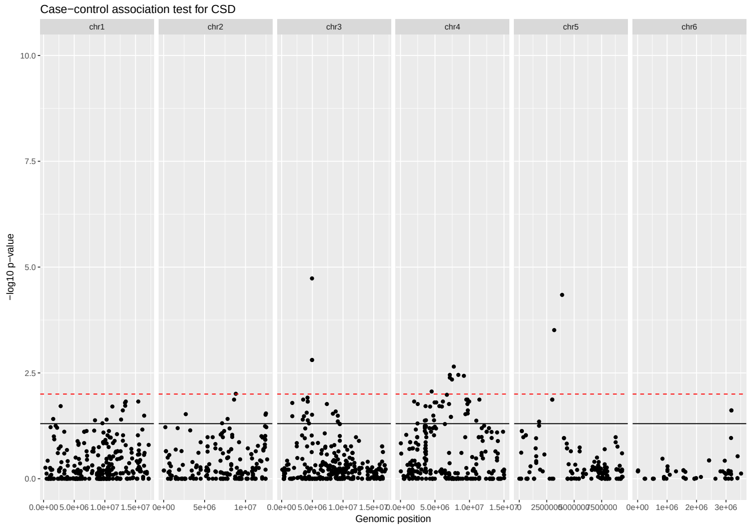

Figure 6: Manhattan plot of the case-control association mapping with fisher exact test. Benjamini-Hochberg correction for multiple testing was applied. Red dashed line: p=0.01, black continuous line: p=0.05

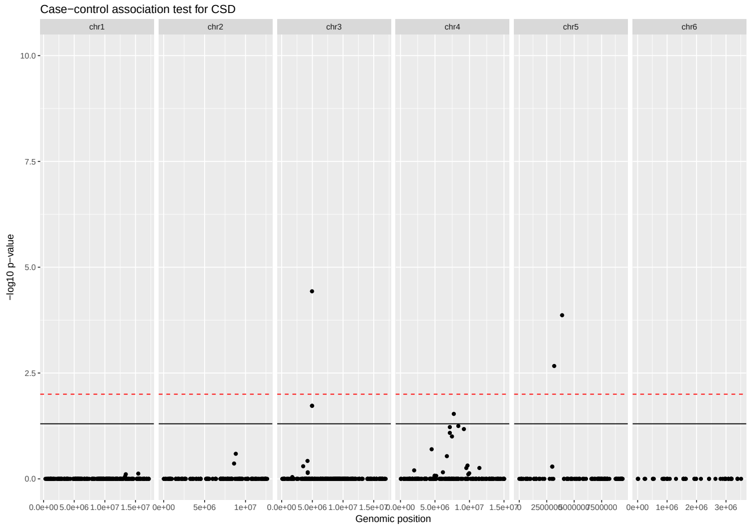

Figure 7: Manhattan plot of the case-control association mapping with fisher exact test. Bonferroni correction for multiple testing was applied. Red dashed line: p=0.01, black continuous line: p=0.05

3. Not sure how to nicely implement centromere into story without mother categories

PS: Figure with proportional size of chromosomes:

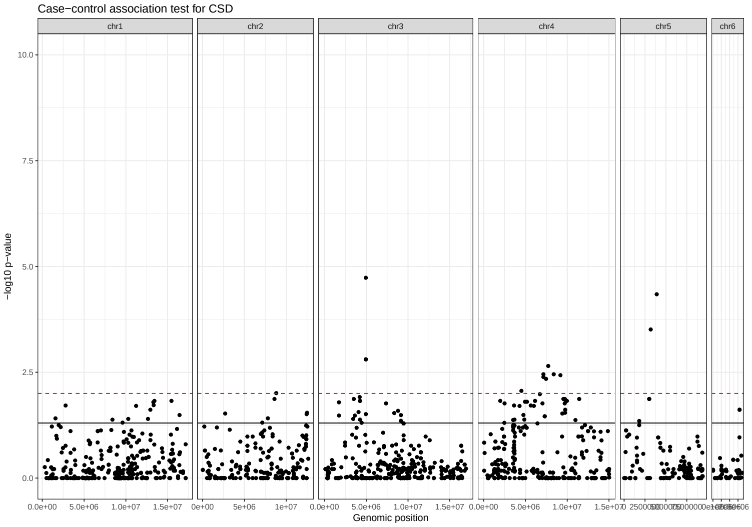

Figure 8: Manhattan plot of the case-control association mapping with fisher exact test. BH correction for multiple testing was applied. Chromosome sizes are proportional. Red dashed line: p=0.01, black continuous line: p=0.05

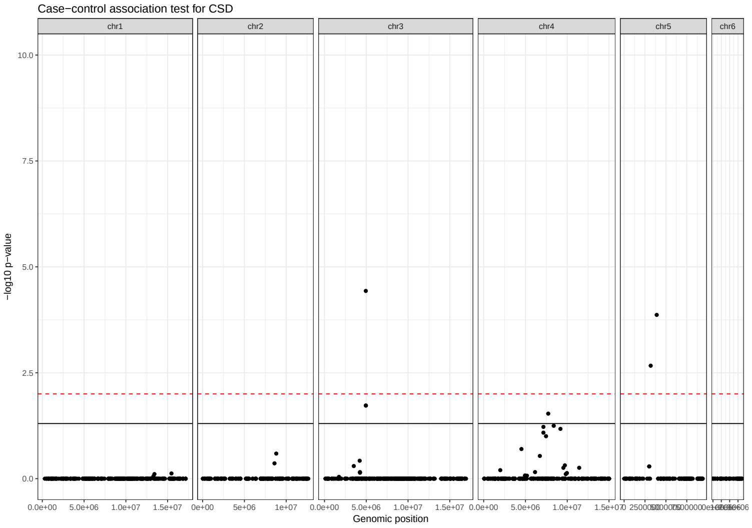

Figure 9: Manhattan plot of the case-control association mapping with fisher exact test. Bonferroni correction for multiple testing was applied. Chromosomes sizes are proportional Red dashed line: p=0.01, black continuous line: p=0.05