# General statistics and progress

*Cyril Matthey-Doret*

*July 24, 2017*

## Questions and ideas

- What would be an expected genome-wide homozygosity value in mothers ?
  - Confirm: STACKS only takes non-consensus loci into account in populations ?
- Technical: When blacklisting loci, is it ok to use Pstacks' snps files of mothers to find the sample IDs of all loci where all positions are homozygous and to look up the catalog ID using their respective sstacks' matches files ?
  - Current method: Removing all snps with no alternative allele in female population displayed (i.e. all females are homozygous) in populations Fst output. Issue: will miss SNPs where mother is homozygous and daughters are heterozygous.
- Looking at the proportion of males among diploid offspring of each mother would allow inferences on the number of heterozygous CSD loci. Once we know these proportion, how can I use this information to improve my power ?
  - Thought: Use list of peaks to find intersect between families with say 2 het. loci and those with 1 het. loci, rather than between all families (i.e. not using intersect between different families with 1 het. locus.)

## Progress and new stuff

- Using fixed threshold identical in all families to separate haploids and diploids (Figure 1), because homozygosity of *haploids* does not depend on mother background.

- Fst probably not so interesting: peaks on averaged plot (Figure 2) caused by a single family (Figure 3), because some SNPs are represented in one or few families.

- After removing loci homozygous in mothers, CSD-prop. plot is looking less noisy (?).

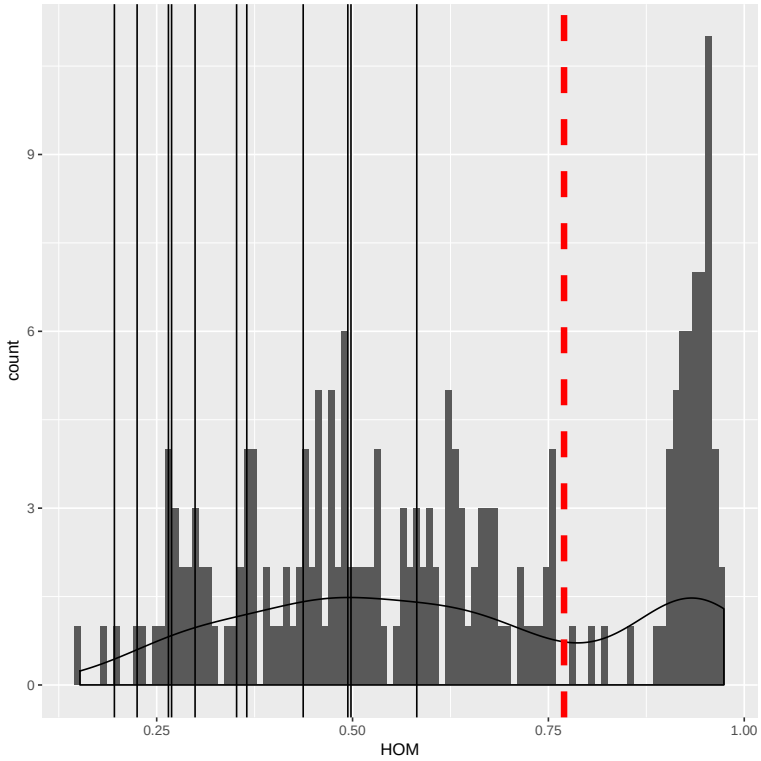



Figure 1: Homozygosity of all samples, including haploids. The red vertical dotted line is the separation threshold. The black continuous vertical lines are the mothers' values.

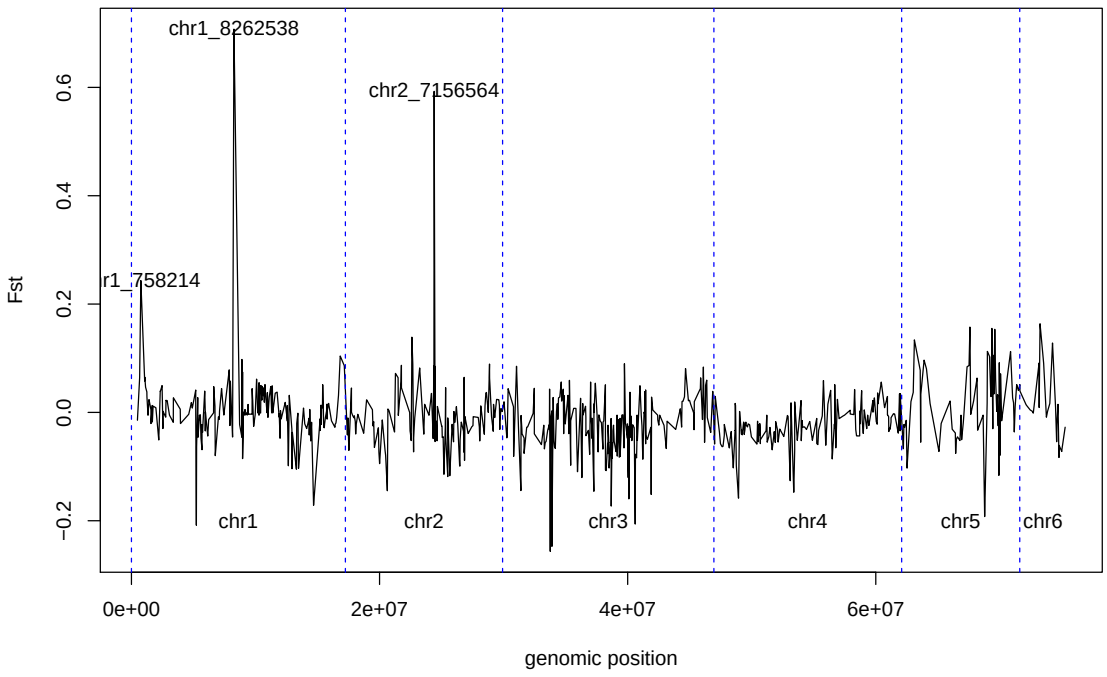



Figure 2: Fst value over genome, averaged across families at each SNPs. Haploids included, did not remove loci homozygous in mothers.


Figure 3: Fst value over genome, overlaying families. Haploids included, did not remove loci homozygous in mothers.

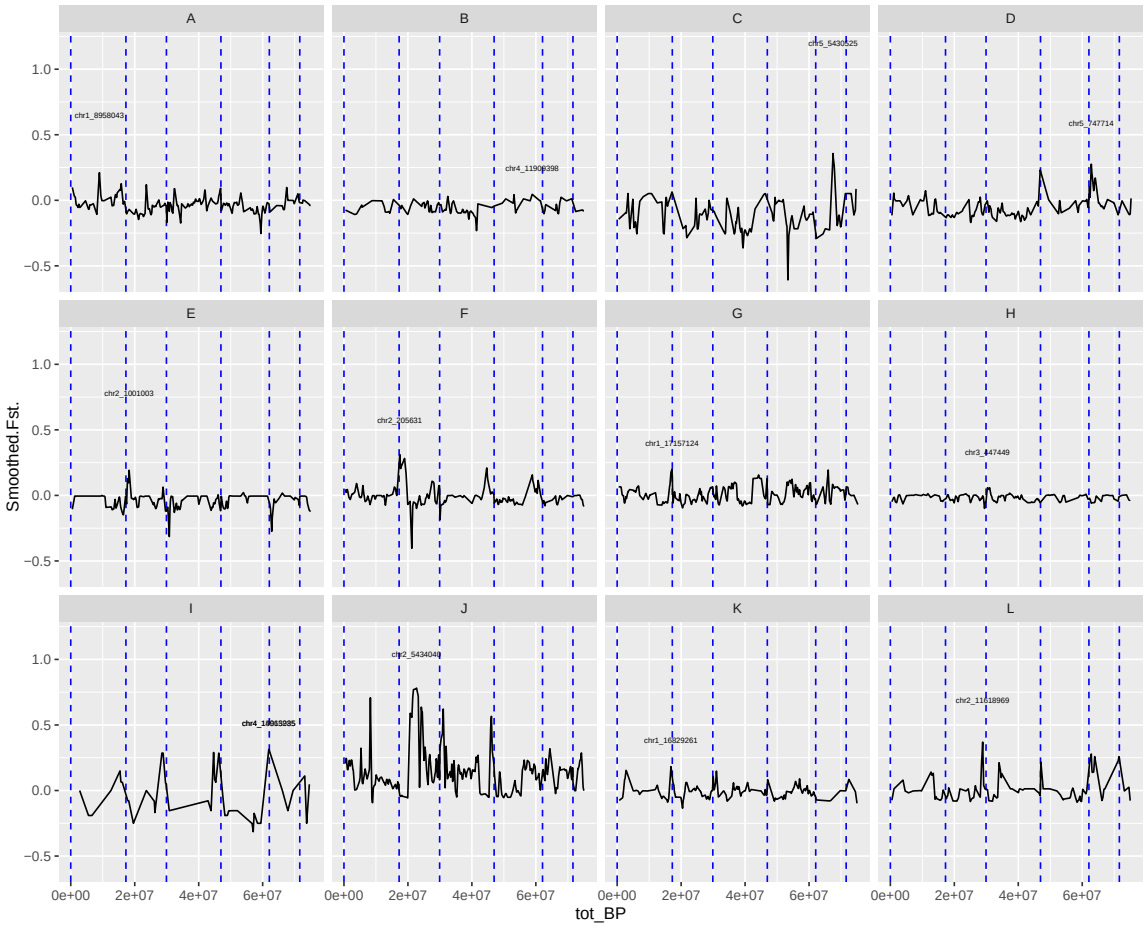
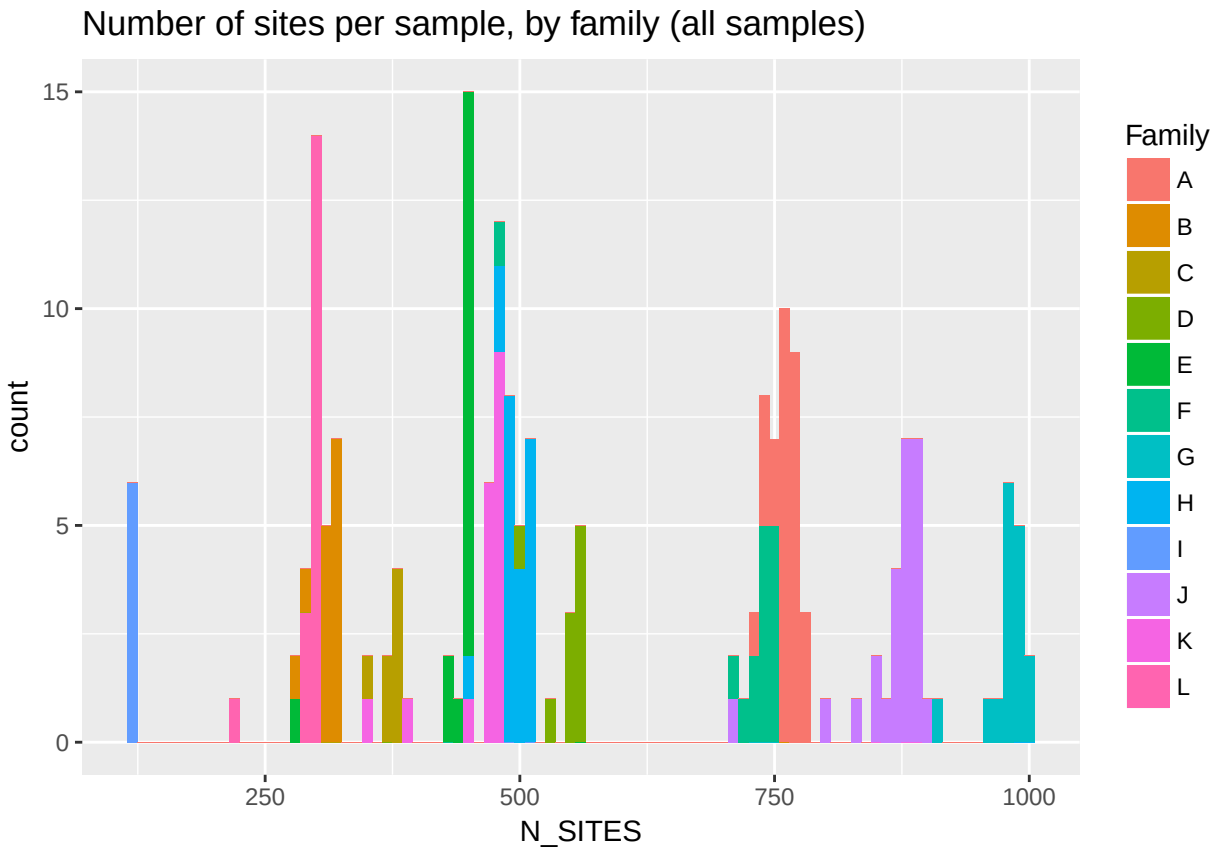



Figure 4: Fst value over genome for each family independently. Haploids included, did not remove loci homozygous in mothers.

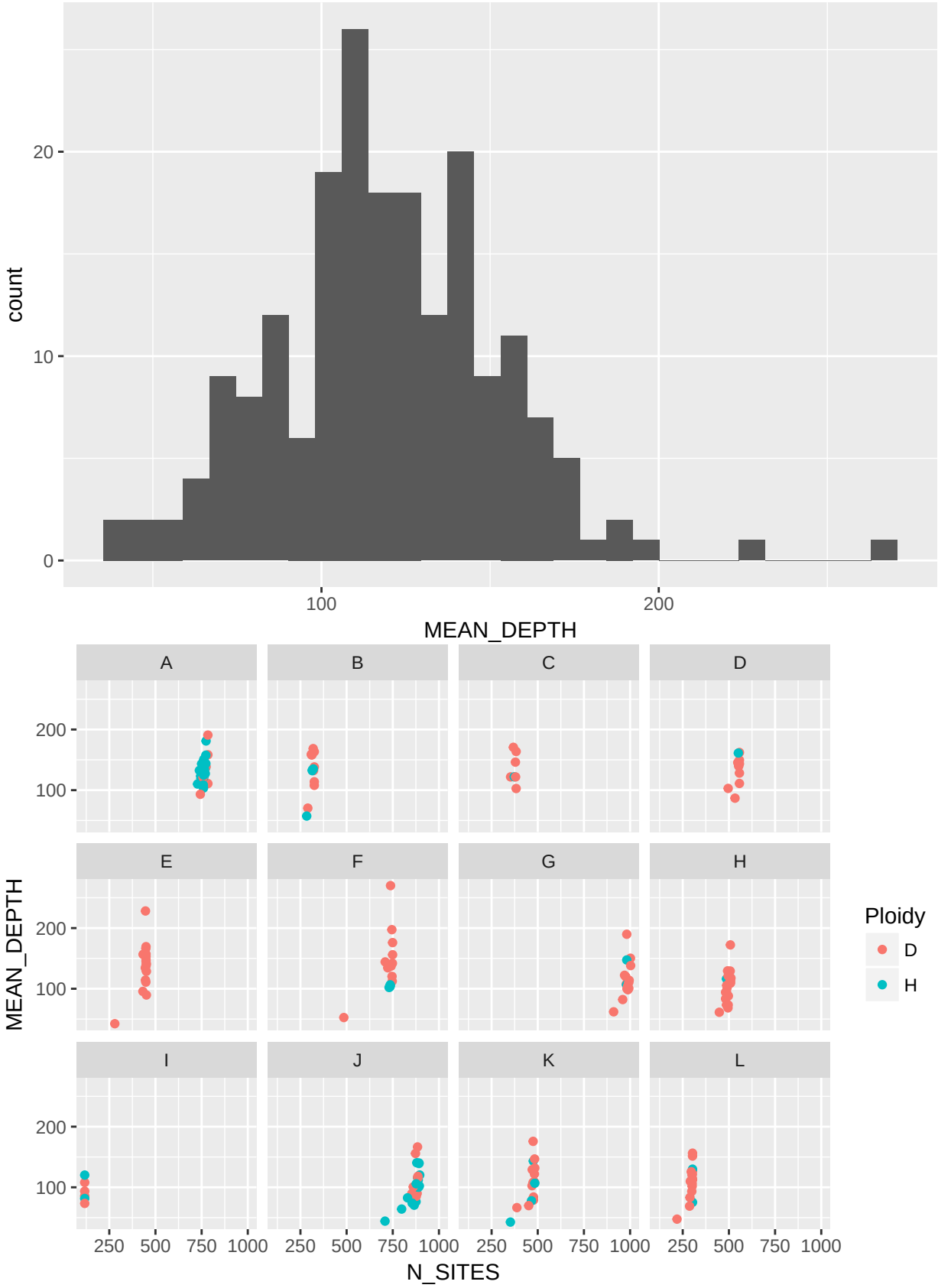
# General statistics

## Number of sites




## Depth

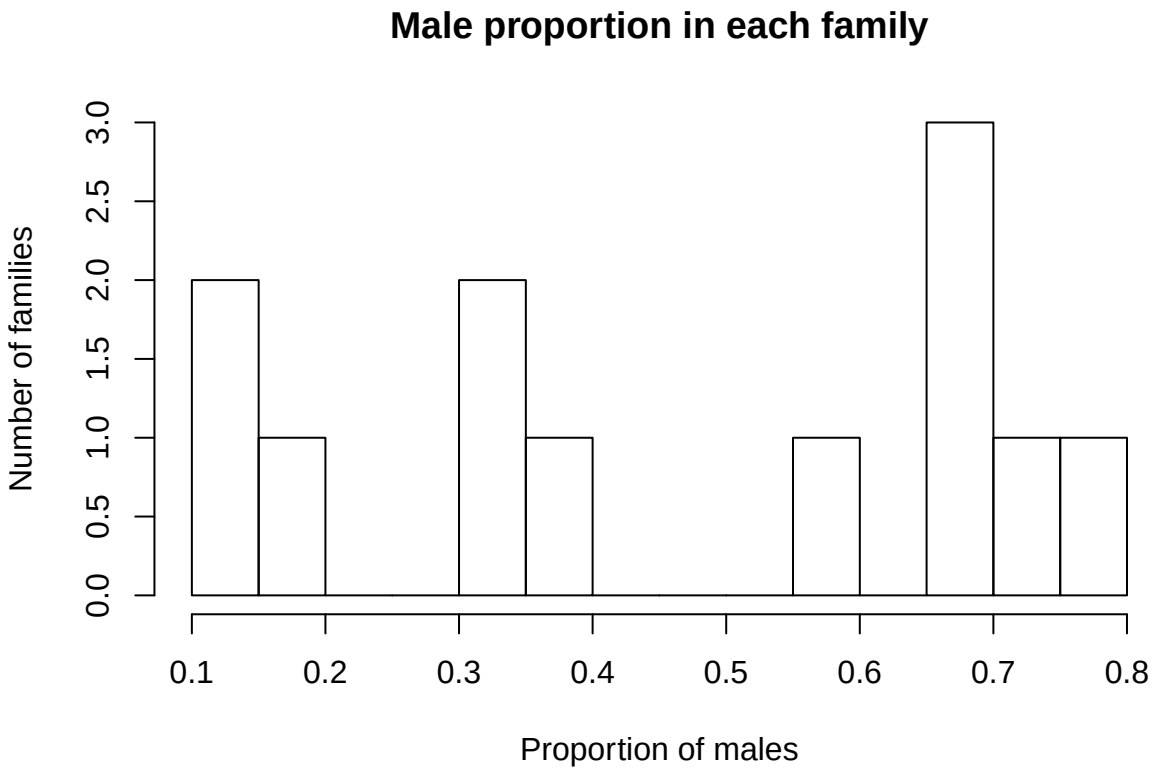
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


count
20
10
0
100
200
MEAN_DEPTH
A
B
C
D
E
F
G
H
I
J
K
L
MEAN_DEPTH
200
100
250
500
750
1000
N_SITES
Ploidy
D
H

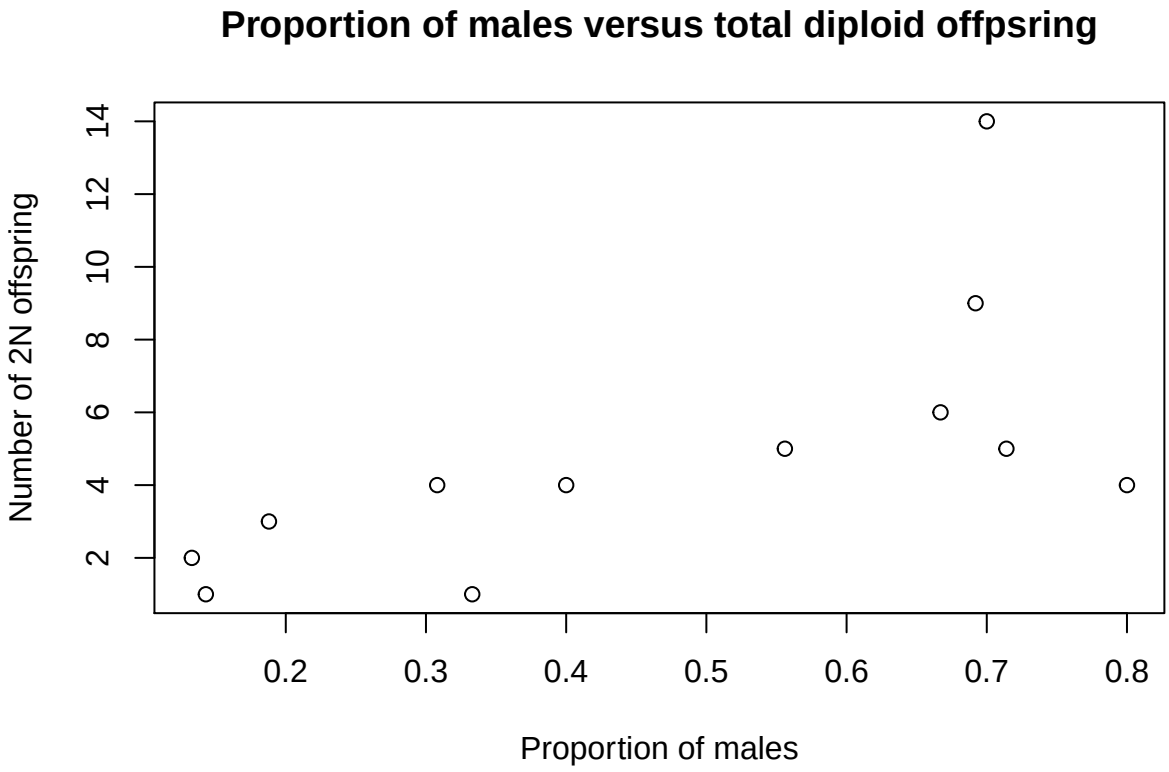
```
## $title
```

```
## [1] "Coverage versus number of sites per individual (all samples)"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

### Offpring proportion




```
## Source: local data frame [12 x 3]
## Groups: Family [12]
##
## # A tibble: 12 x 3
##    Family Count prop_males
##    <fctr> <int>      <dbl>
##  1      A     5      0.714
##  2      B     6      0.667
##  3      C     4      0.800
##  4      D     5      0.556
##  5      E     3      0.188
##  6      F     4      0.400
##  7      G     4      0.308
##  8      H    14      0.700
##  9      I     1      0.333
## 10      J     1      0.143
## 11      K     9      0.692
## 12      L     2      0.133
```




## Coverage across genome

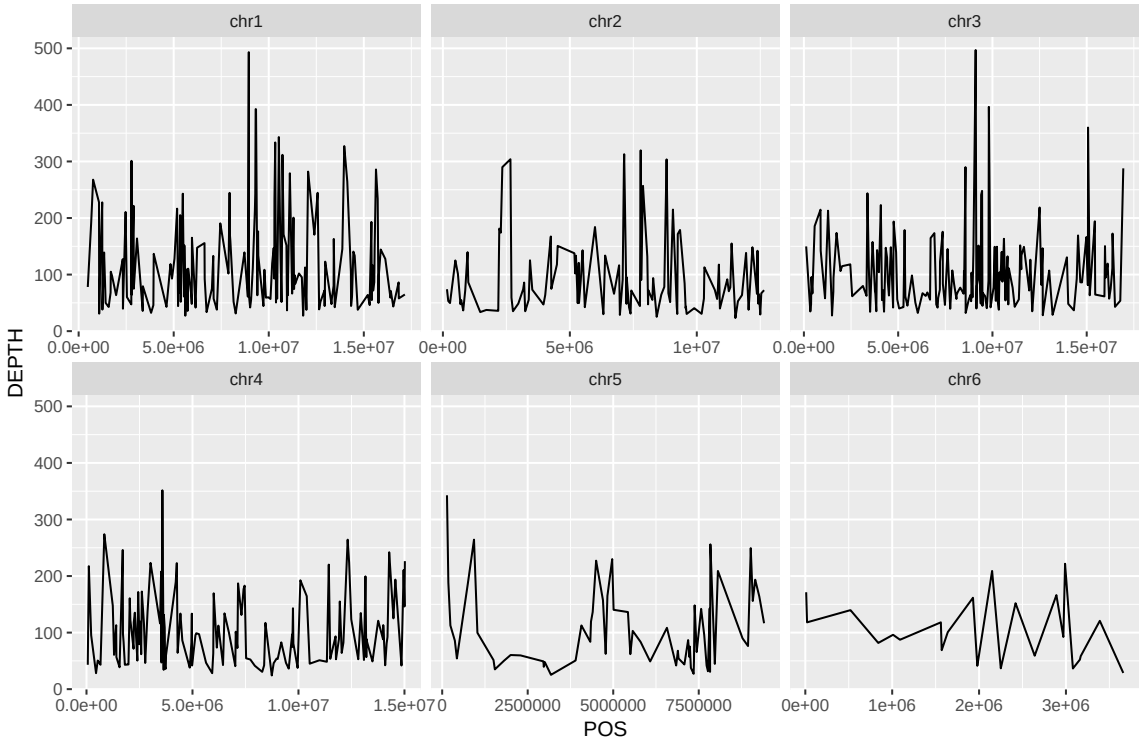



Figure 5: Coverage across genome, averaged across all individuals by 10kb windows.

### Last minute plots

```
# Proportion of offspring homozygous at loci heterozygous in mother
```