Genotype inference analysis

Ruth Gómez Graciani

In this report, I make descriptive plots and statistical analyses to find whether inversion regions have a tendency to have lower recombination rates in heterozygous individuals.

1 Origin of the data

1.1 Genotypes

Genotypes were imputed in our 20 individuals using IMPUTE2, tagSNP inference or both. Genotypes' quality control report can be found in "report/2020-10-28_genotypeFilteringReport/filteringAnalysis.pdf". We obtained more than 3 high-quality genotypes coming from both homozygous and heterozygous individuals for 60 inversions.

1.2 Map

Recombination maps were calculated from recombination events in a probabilistic way. The genome is divided into windows, for which recombination rates are calculated following a probabilistic method: instead of just assuming that the crossover took place in the center of the recombination event, each event is ponderated depending on how much of it is overlapping with a window, and the sum is used to calculate cM/Mb values for each window. Then, recombination results are normalized using a quantile normalization in order to make them comparable.

The effectivity of this method, as well as the smallest informative window size, were assessed with simulations. For each recombination event, a hypothetical actual location for the crossover was randomly selected and then the corresponding recombination rate calculated. We obtained the correlation between the simulated rates and the rates calculated with low-resolution recombination events. This gives us a measurement of how close estimated rates would be to real ones. The probabilistic method proved to be better than the center-point method, and window sizes between 150 and 200kb (corresponding to 0.9 and 0.95 correlations) would be optimal (Figure 1).

2 Descriptive analysis

2.1 Inversion groups

It could be that inversions with different sizes show different behaviors. To account for that, I want to divide inversions in two groups of size. There is a bias in the distribution of genotyped inversions when compared with all the available inversions because small ones tend to be NH-generated and unique and big ones to be NAHR-generated and thus probably recurrent, making small inversions more likely to be correctly genotyped (Figure 2). Thus, I decided to calculate the classification thresholds from the original distribution rather than the genotyped inversions distribution, which would return skewed thresholds.

I tested two classification thresholds, the median and the 3rd quantile, rounded to 9kb and 24kb. Table 1) shows some basic information about the groups

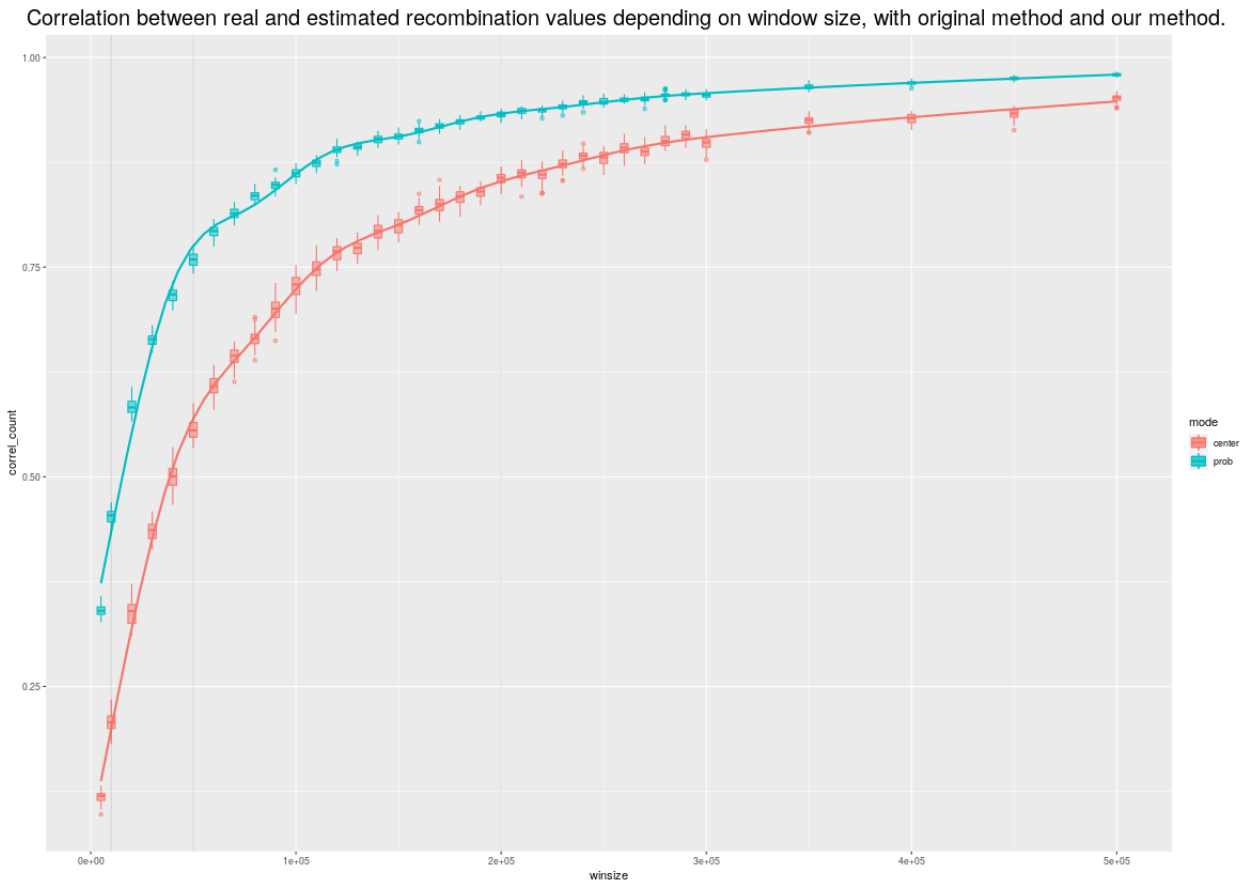

Figure 1: Real recombination rates were simulated at different window sizes and comapred with the corresponding estimated ones. According to this result, our probabilistic method is more accurate than the center-point method, and the minimum informative window size is 150-200 kb.

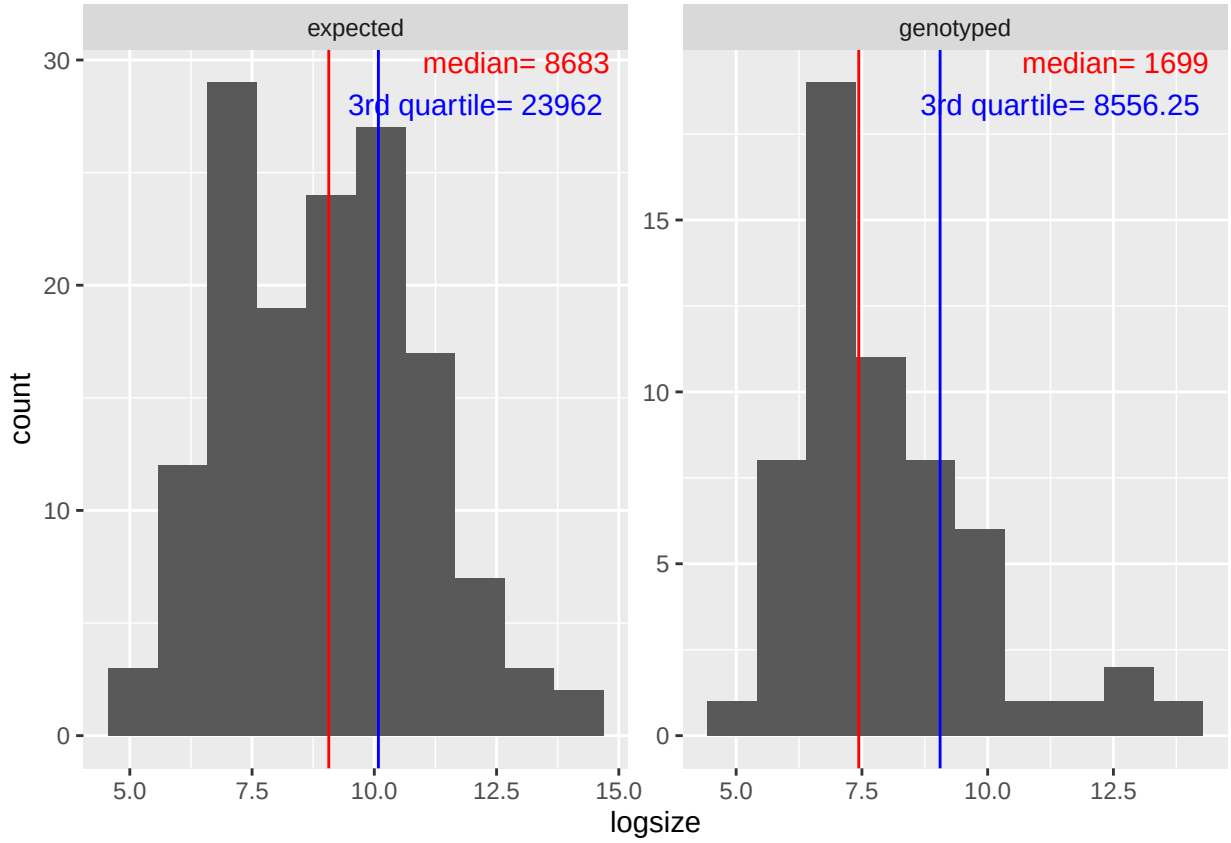

Figure 2: Size distributions (log transformed) for all the available inversions (expected) and for the actually genotyped inversions. Small inversions are genotyped proportionally to the original distribution while big inversions are less often correctly genotyped, probably because they tend to be NAHR-mediated, recurrent inversions. This changes in distribution generates a big bias in the median and 3rd quantile, so the thresholds were calculated using the expected distribution.

Table 1: Basic information for the two group sizes in which inversions were divided

	Median			3rd Quantile		
Group	n.Invs	minsize	maxsize	n.Invs	minsize	maxsize
Big	14	9329	939248	5	42967	939248
Small	44	133	8717	53	133	19311

2.2 Recombination rate differences

The selected window size is bigger than most inversions. In those cases that the inversion spans more than one window, we made the mean of the windows. Then, for each inversion, the mean values for heterozygous and homozygous individuals was calculated and to compared both gorups with the fold change (log2(Homozygous/Heterozygous)).

HsInv0325 was considered an outlier for now.

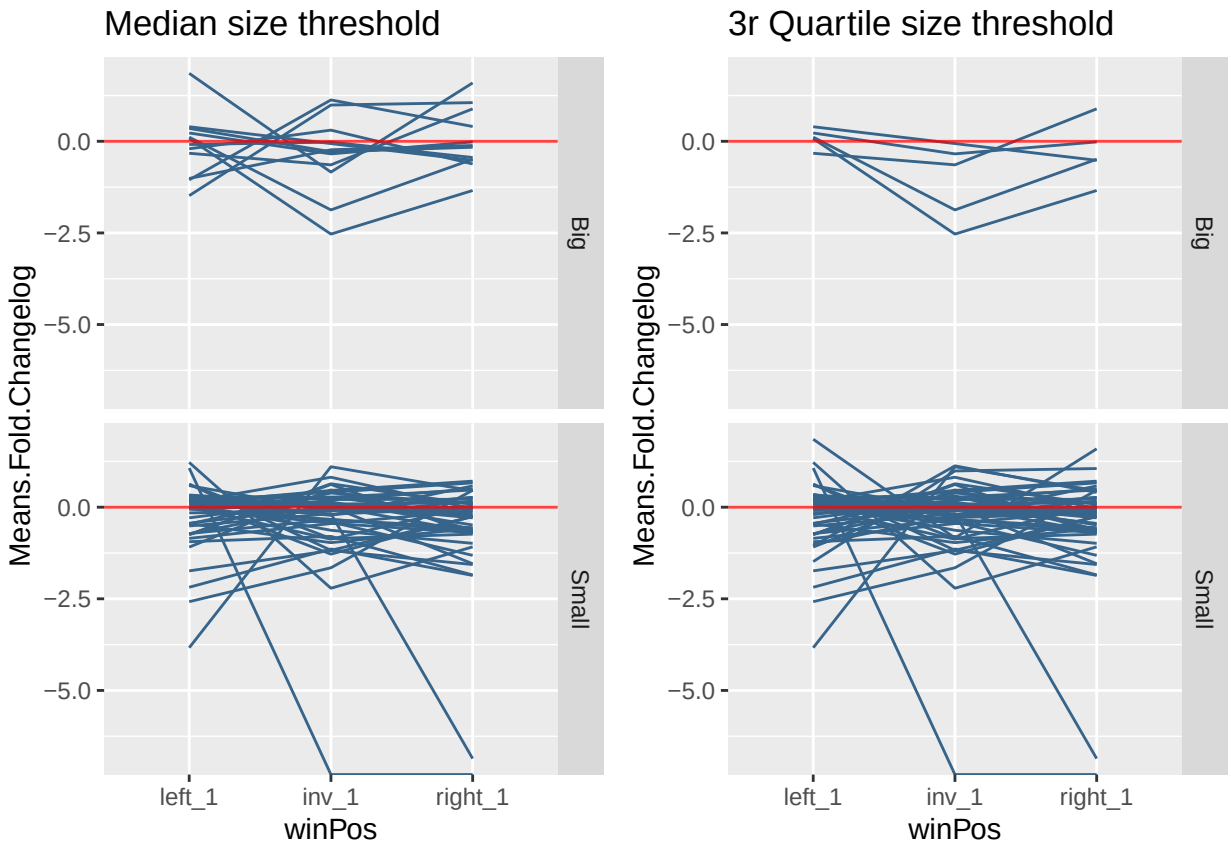

Figure 3: Line plots showing tendencies for big and small inversions, using two thresholds to define the categories. HsInv0325 was considered an outlier and removed

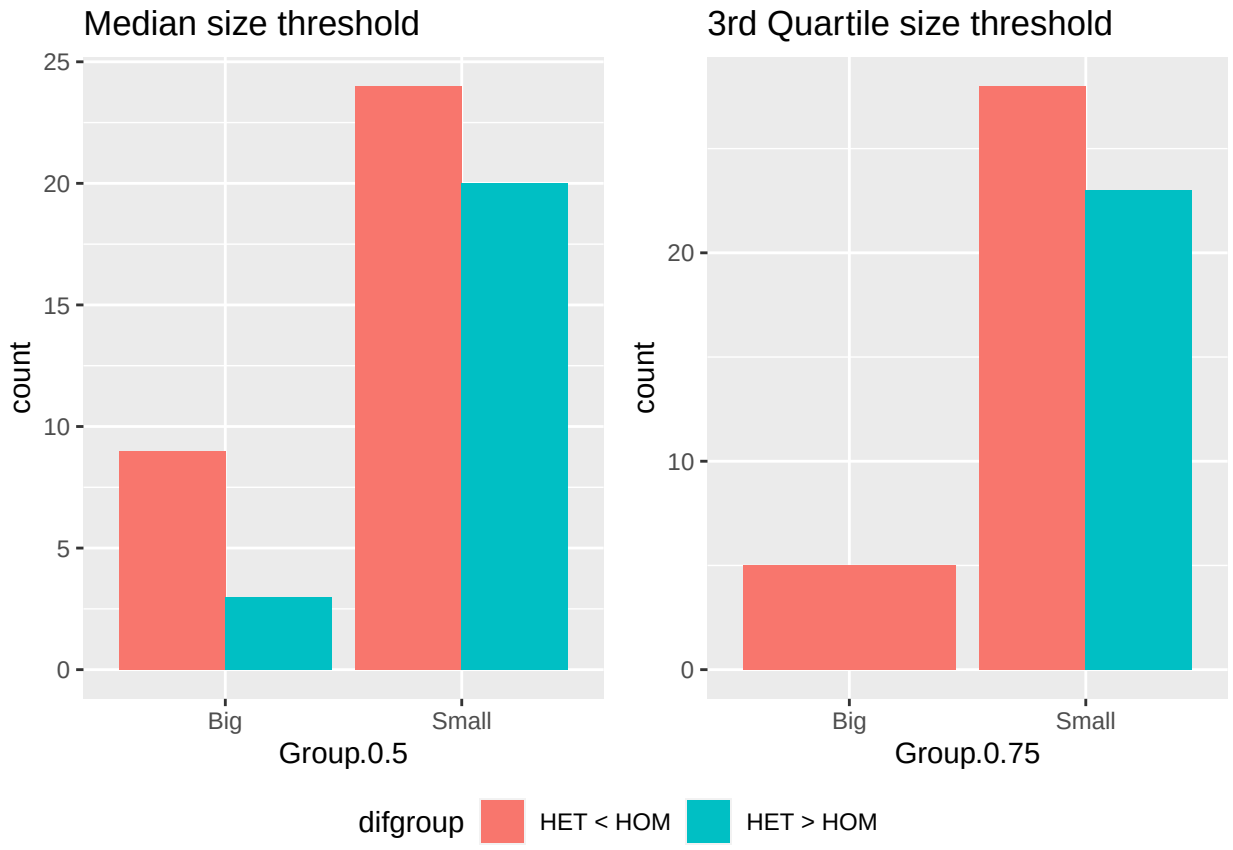

Figure 4: Amount of cases with Heterozygous > Homozygous and viceversa.

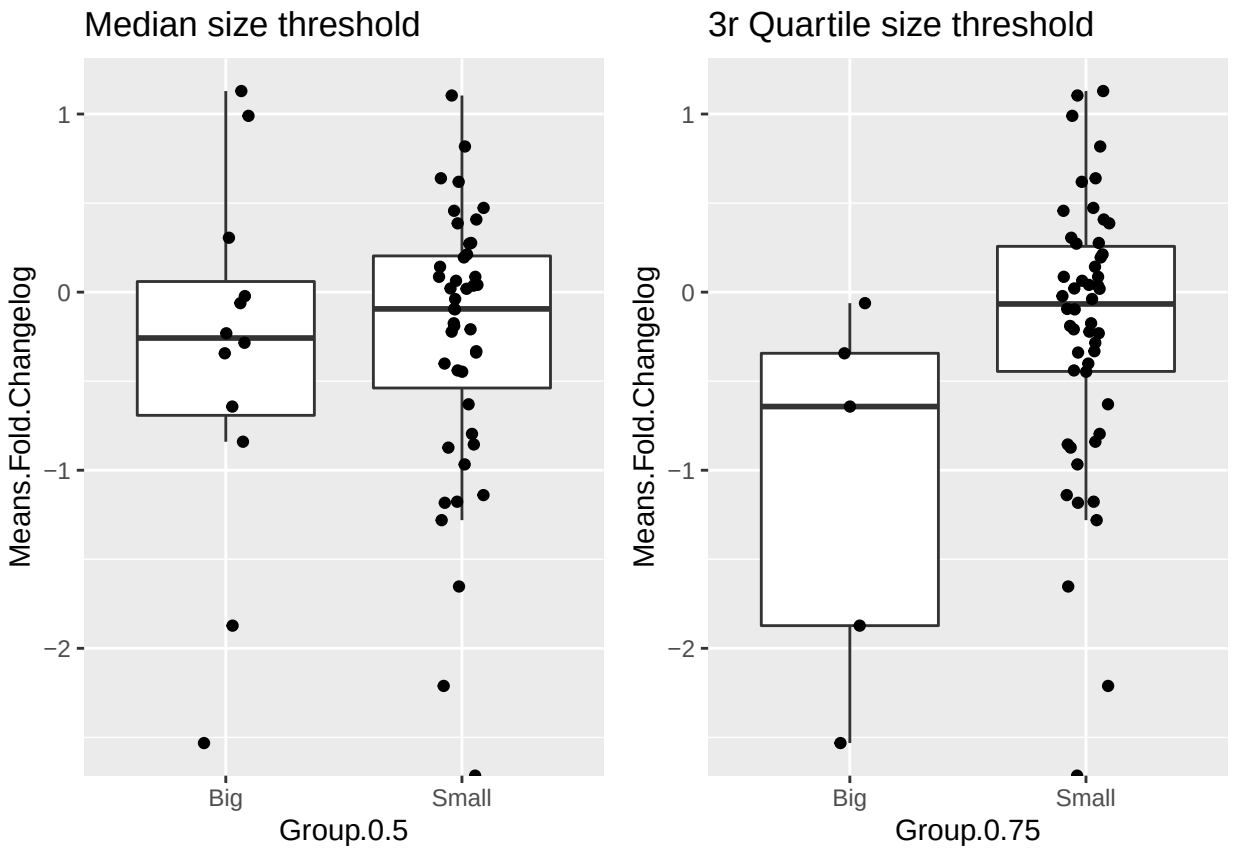

Figure 5: Box plots showing tendencies for big and small inversions, using two thresholds to define the categories. HsInv0325 was considered an outlier and removed

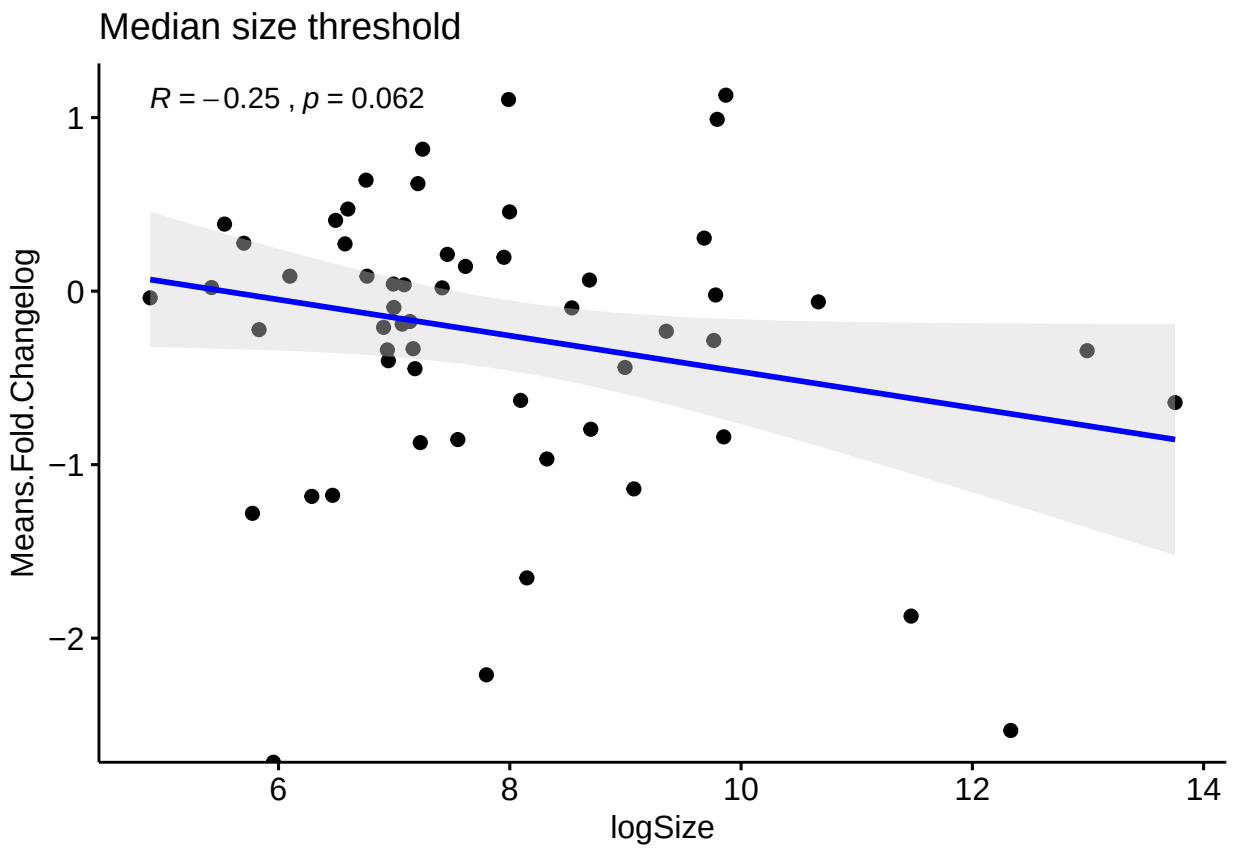

Figure 6: Correlation between log transformed size and fold change after removing outliers.

3 Statistical analysis

```
## [1] "Chi squared test using Median threshold"
```

```
##
##             Big Small
##   HET < HOM   9    24
##   HET > HOM   3    20
```

```
## Warning in chisq.test(Matriz, correct = TRUE): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Matriz
## X-squared = 0.89432, df = 1, p-value = 0.3443
```

```
## [1] "-------------------------------------------------"
```

```
## [1] "Chi squared test using 3rd Quartile threshold"
```

```
##
##             Big Small
##   HET < HOM   5    28
##   HET > HOM   0    23
```

```
## Warning in chisq.test(Matriz, correct = TRUE): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Matriz
## X-squared = 2.19, df = 1, p-value = 0.1389
```