Genotype inference analysis

Ruth Gómez Graciani

In this report, I analyze the differences in local recombination rates between heterozygous and homozygous individuals for an inversion.

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 61 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).

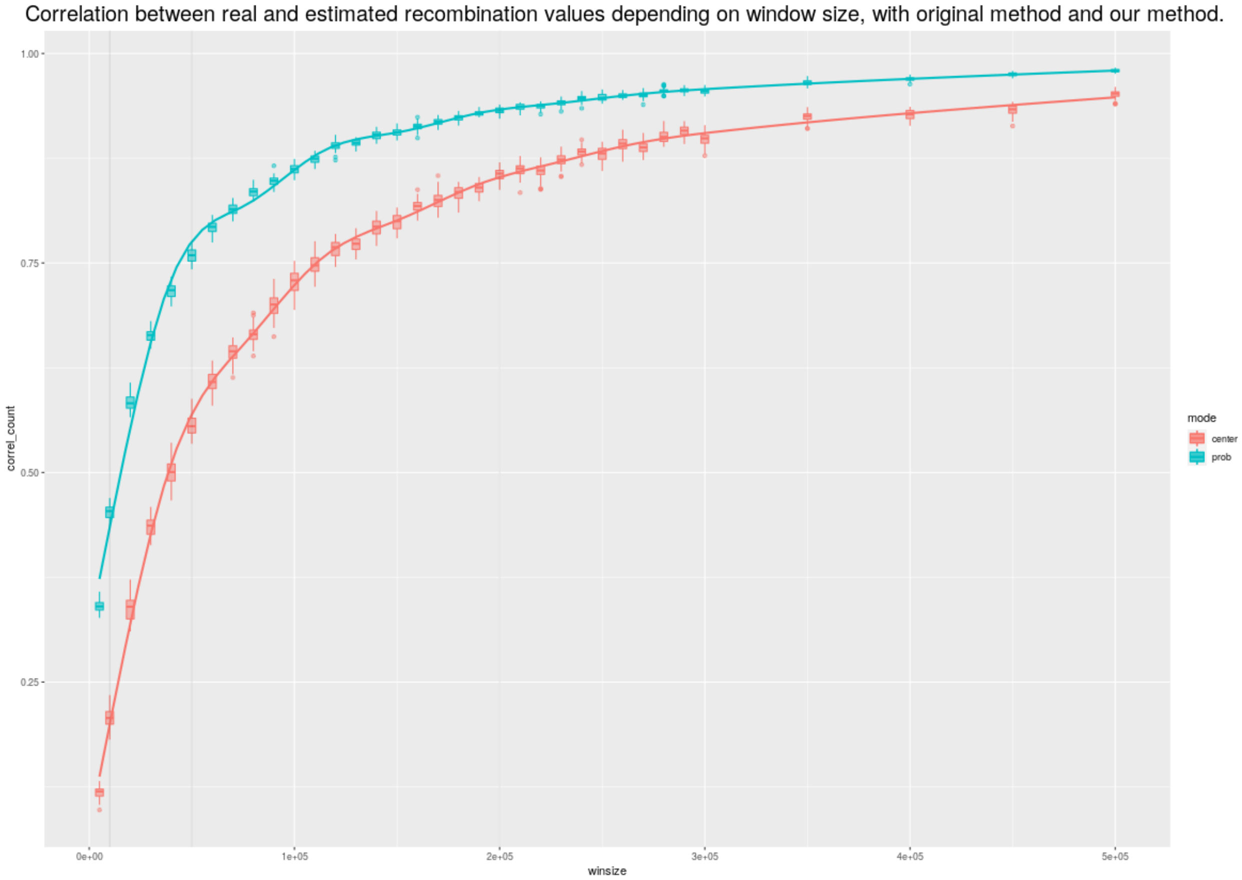
2 Figures

Include only the most relevant figures

– For later

Appendix: all the tested figures

This appendix includes all the tested figures to avoid having to repeat them or not remembering if we tested something.



Correlation between real and estimated recombination values depending on window size, with original method and our method.

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.

3 Data selection

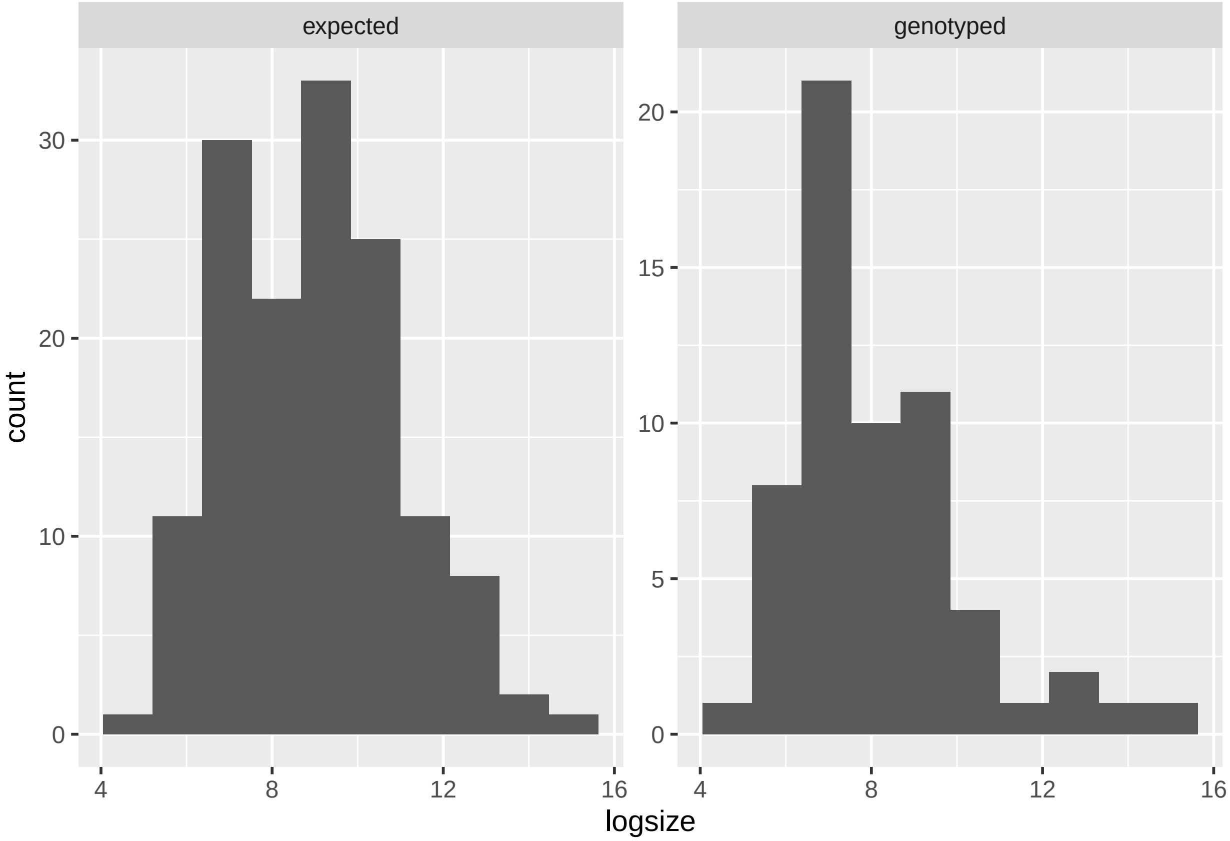
Here I list all the variables and data trimming criteria that were at some point considered or suggested.

- **Nominal variables:**
  - Breakpoint type (simple/complex)
  - Presence of inverted repeats
  - Breakpoint definition (Very well characterized and reliable vs. Not sure)
  - Physical length (natural or transformed to a normal distribution)
  - Genetic length (natural or transformed to a normal distribution)
- **Measurement variables :** 1 value per inversion - and window, if analyzing more than one window.
  - Fold change of the Normalized Recombination Rate between heterozygous and homozygous individual means
  - P-value of the comparison between heterozygous and homozygous means through a Student's T test
  - Power of the aforementioned Student's T test
- **Data trimming criteria:** aka, data to remove.
  - Outliers
  - Inversions <1kb

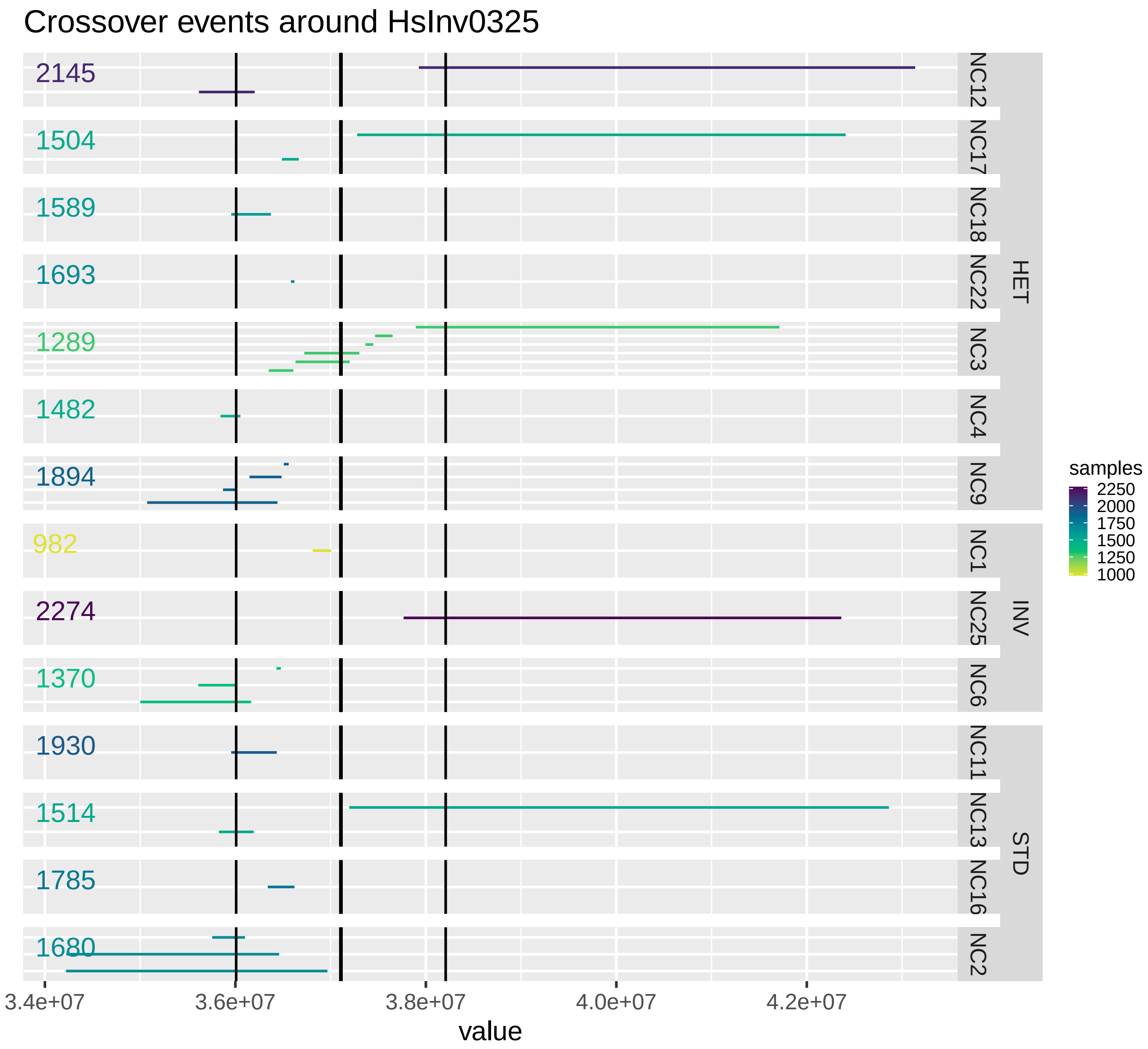
4 Raw measurements visualization

4.1 Size bias in genotyped inversions.

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.



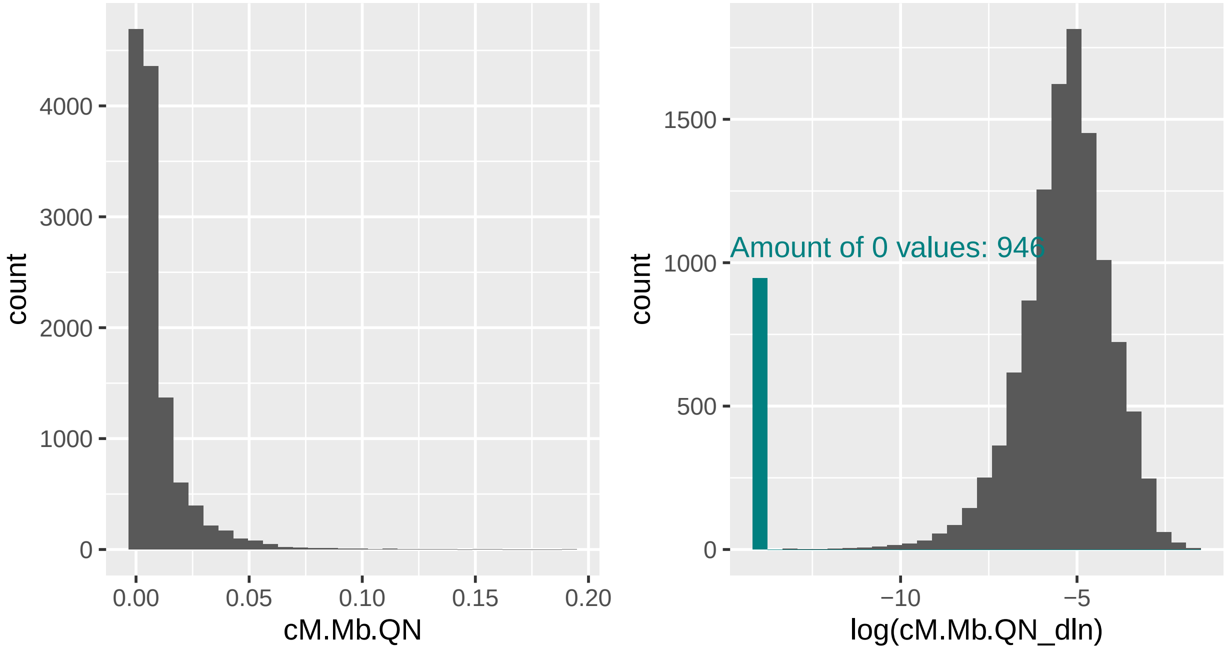
4.2 Crossover events overlapping with the studied region around genotyped inversions



4.3 Quantile-Normalized Recombination Rate values distribution in genotyped inversions

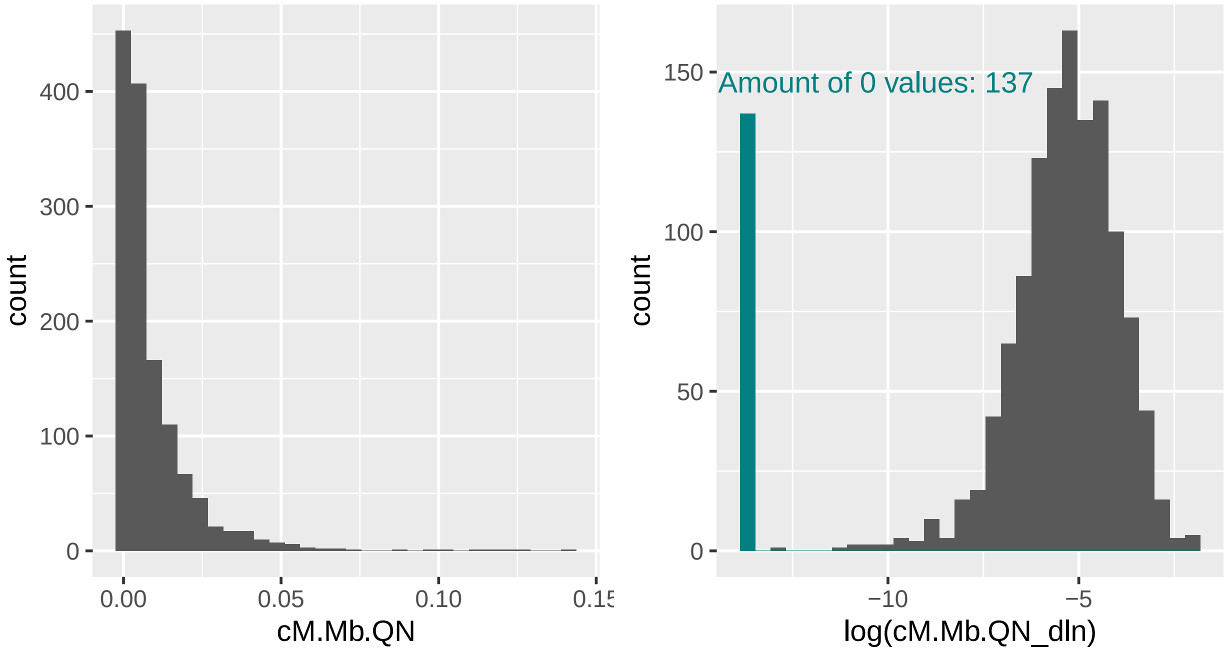
4.3.1 General distribution

All windows and inversions.



4.3.2 Inside inversion distribution

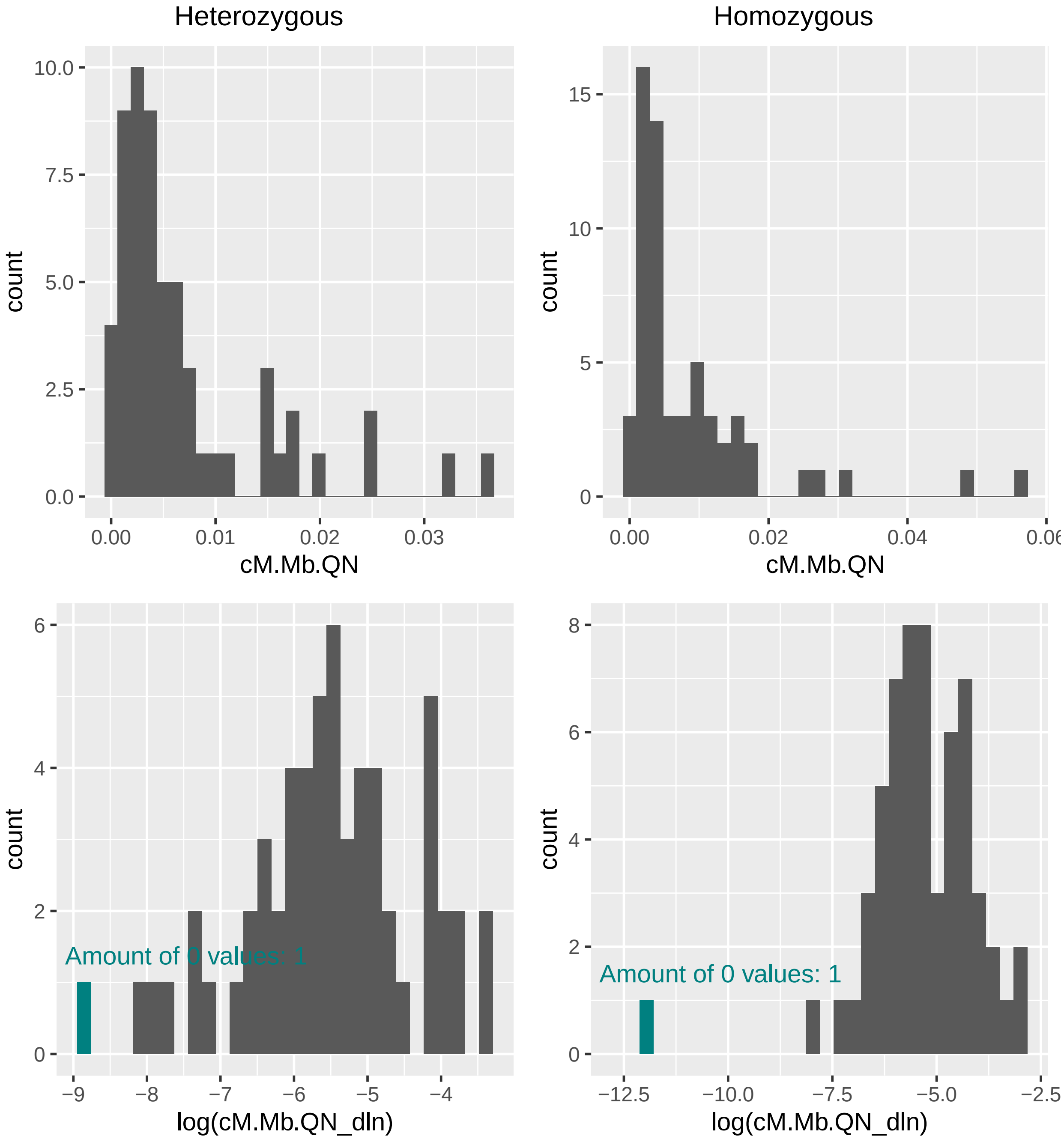
All windows from inside inversions (some inversions have more windows than others).



4.3.3 Mean inside inversion windows distribution

The mean value for the windows inside inversions (one window per iversion and individual).

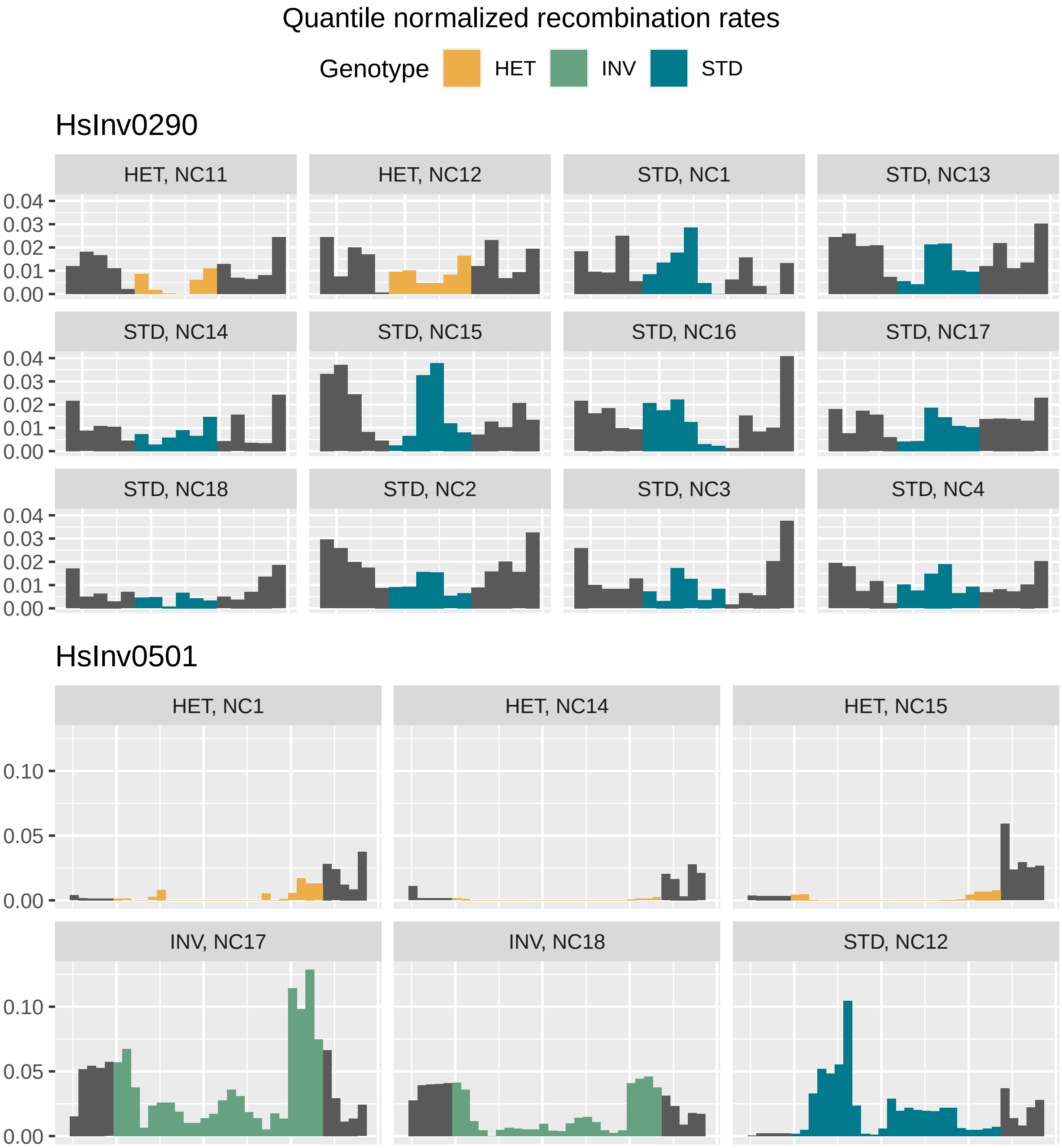
To make this plot better, make scales for Heterozygous and homozygous equal! Note that heterozygous individuals have lower values in general



4.4 Local recombination rate panoramic visualization in genotyped inversions

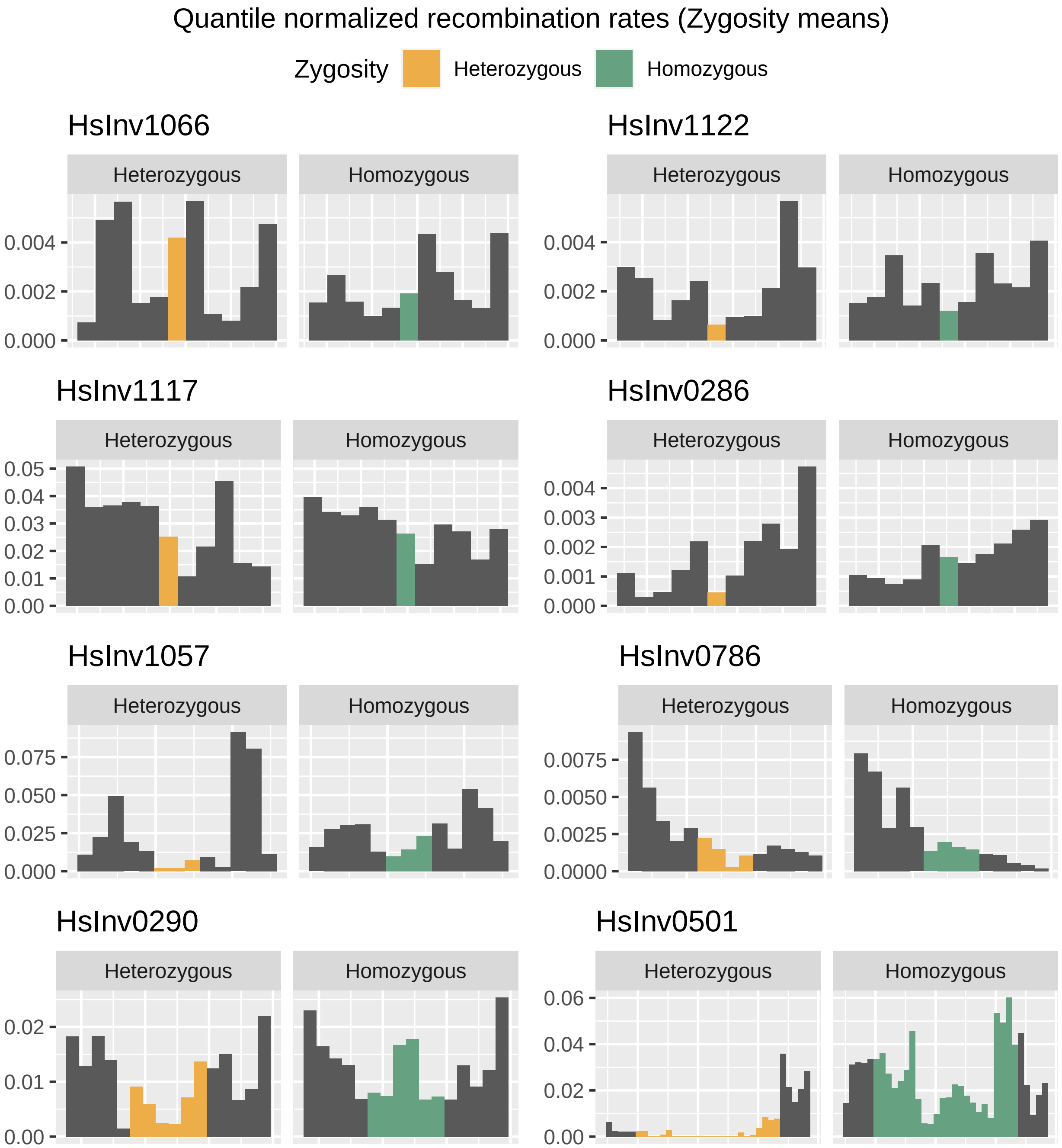
4.4.1 All data

This is only a sample, but a figure with all the inversions or only a selected group can be generated



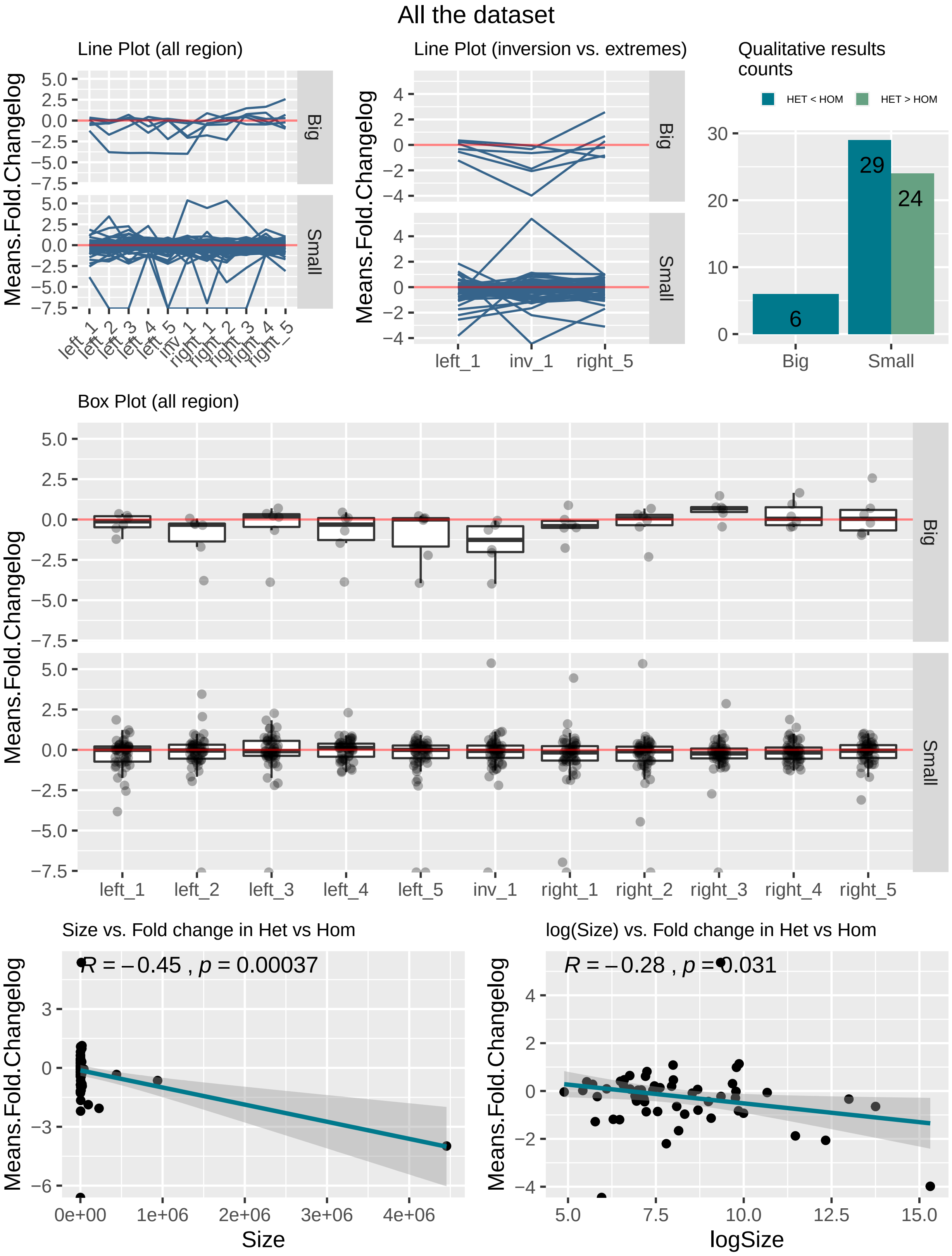
4.4.2 Homozygous vs. Heterozygous means

This is only a sample, but a figure with all the inversions or only a selected group can be generated.

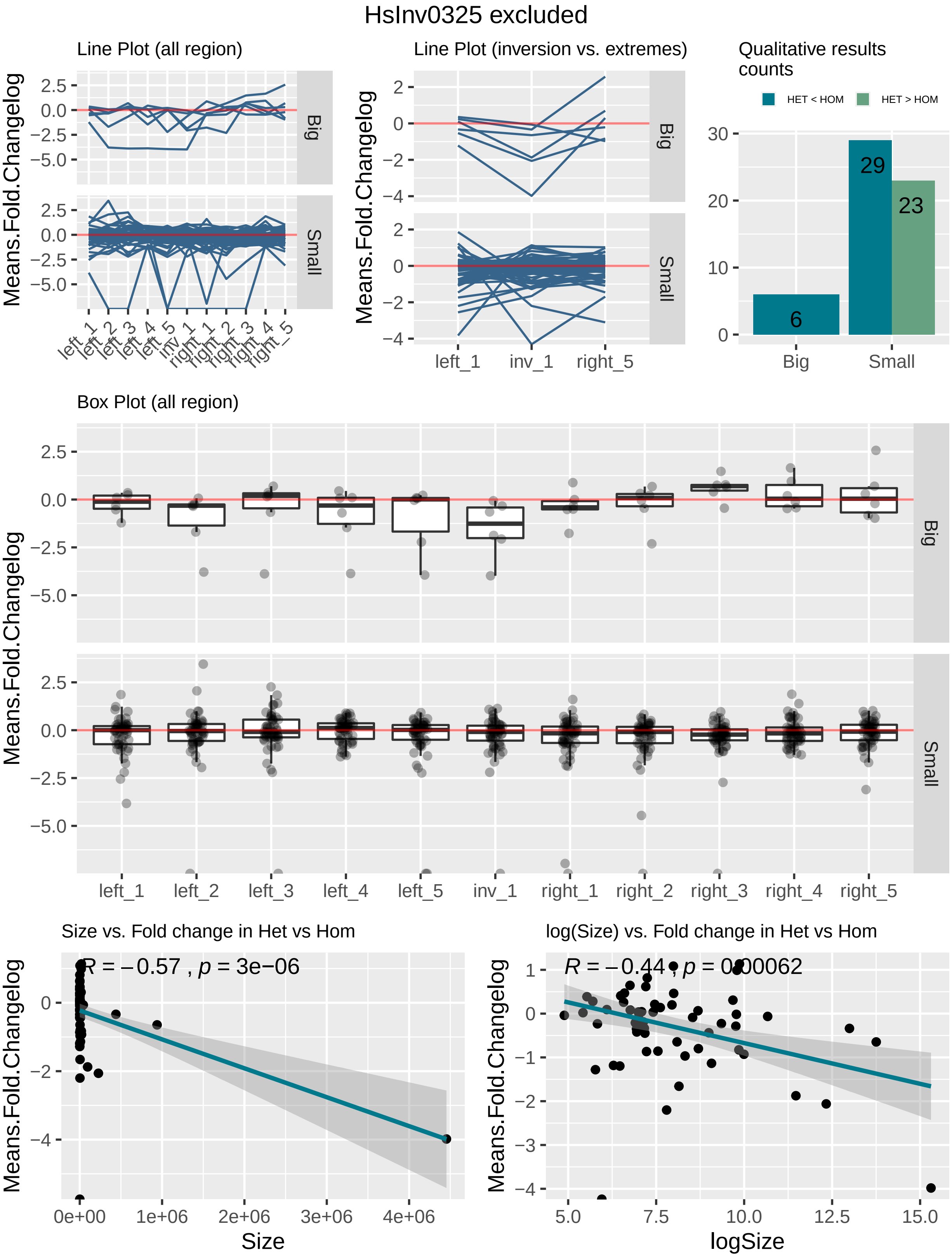


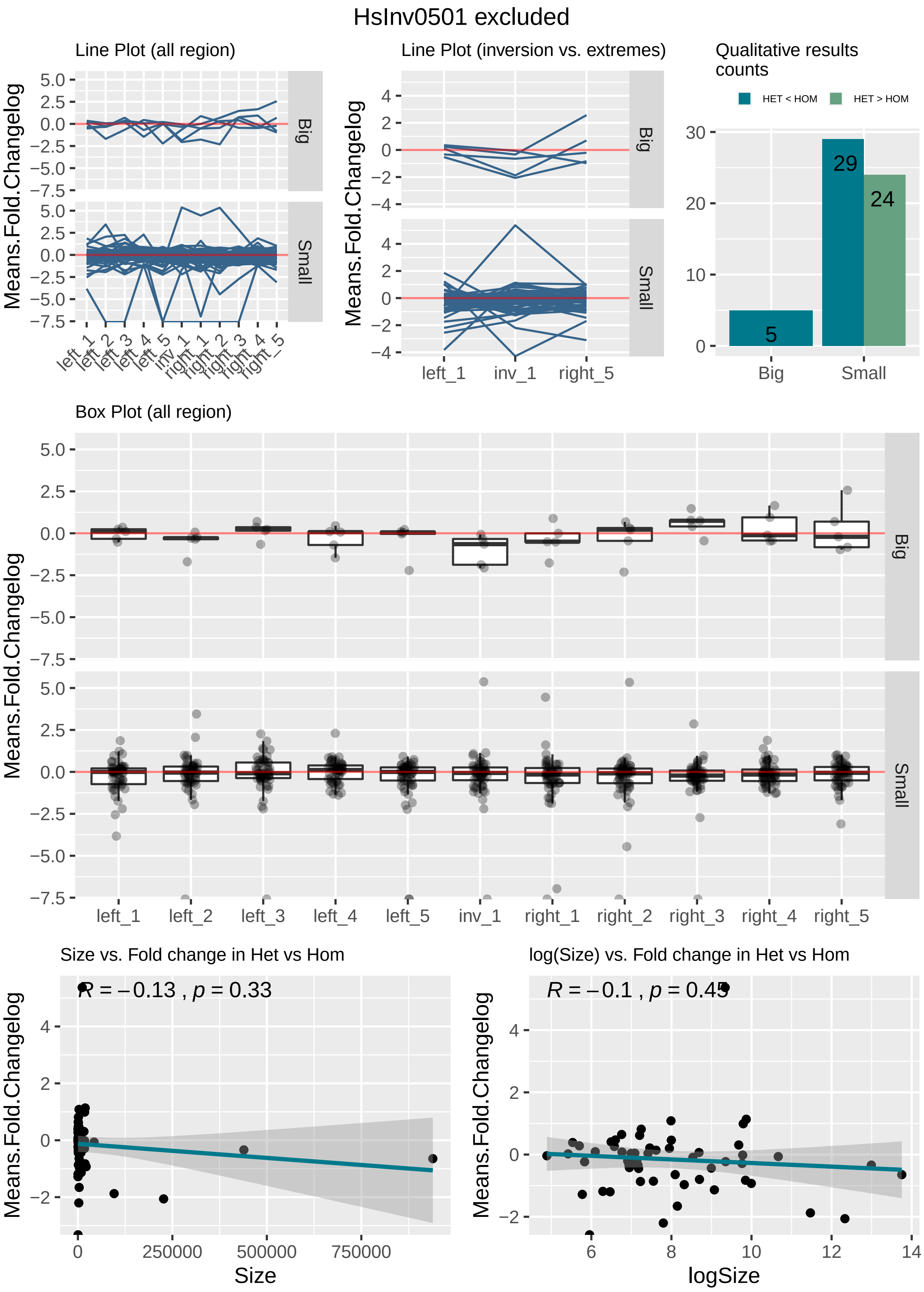
5 Heterozygous vs. Homozygous formal comparison

5.1 All the dataset

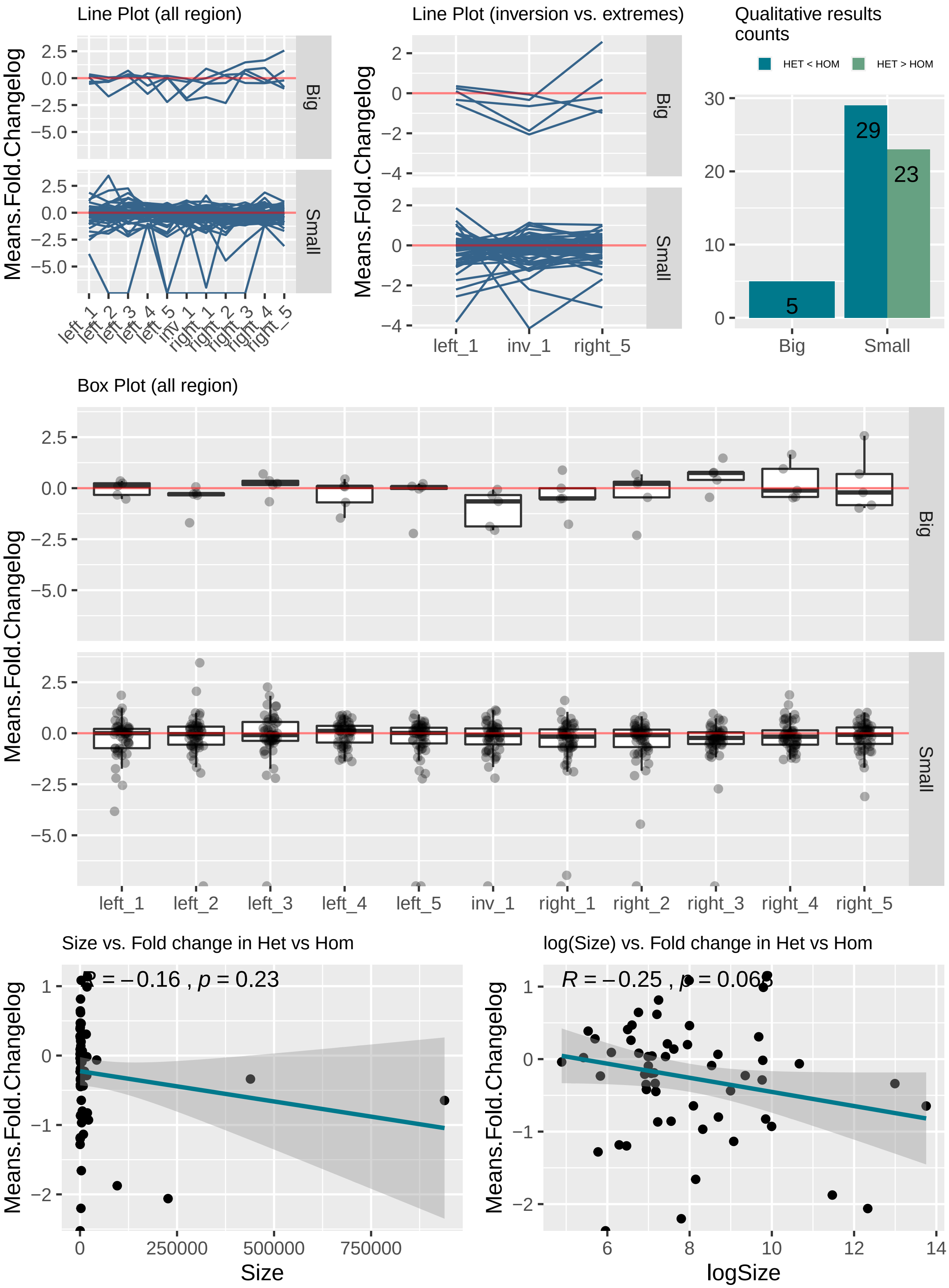


5.2 Dataset without outliers




HsInv0501 excluded
Line Plot (all region)
Means.Fold.Changelog
5.0
2.5
0.0
−2.5
−5.0
−7.5
Big
Small
left_1
left_2
left_3
left_4
left_5
inv_1
right_1
right_2
right_3
right_4
right_5
Line Plot (inversion vs. extremes)
4
2
0
−2
−4
Qualitative results counts
HET < HOM
HET > HOM
30
20
10
0
5
29
24
Box Plot (all region)
Size vs. Fold change in Het vs Hom
R = −0.13 , p = 0.33
0
250000
500000
750000
Size
log(Size) vs. Fold change in Het vs Hom
R = −0.1 , p = 0.45
6
8
10
12
14
logSize

HsInv0325 and HsInv0501 excluded



6 Modelo nulo con genotipos al azar

7 HET vs HOMO in Student's T p-value. 1 point = 1 inv

- Boxplot per finestres - left in right, left i right haurien de ser mes iguals que in
- Correlació mida física vs pval
- Correlacio mida genètica vs pval

8 HET vs HOMO in power calculations. 1 point = 1 inv

9 Chi square - proporció de positius i negatius

- Plot de positius i negatius
- Chi square de les proporcions entre positius i negatius, separats per mida

10 Outliers

- Outlier selection
- Outlier explanation
- Solapament entre events de recombinatió (raw data) i inversions.
- Plot de panorama (normalized recombination rate) per 1 sola inversio.
- Individu a individu
- Mitjana HET vs HOMO
- p-value of the normalized rec.rate in the out distribution, to compare between invs, especially for outliers?

11 ——————————————————

–>

–> –> –> –>