

Inversions position depending on chromosome characteristics

Location patterns 1

Ruth Gómez Graciani

January 11, 2022

1 Introduction

Inversions can be classified in two groups depending on their generation process: two relatively close breaks that are repaired in opposite orientations by non-homologous mechanisms (which we call NH inversions), or non-allelic homologous recombination between inverted repeats (NAHR inversions). These two groups have differences regarding important characteristics such as size, location and divergence between orientations. NH inversions are smaller, can appear anywhere in the genome and are unique, which causes a marked divergence between opposite orientations as time goes by. NAHR inversions appear under the specific conditions of having inverted repeats at a certain distance, but can be much bigger than NH inversions and can be generated more than once throughout the history of a population, which leads to less divergence between orientations [1][2].

In addition, inversions generate aberrant chromosomes and unviable zygotes if a recombination event takes place within the region during meiosis. Thus, they are expected to have better chances of reaching polymorphism in low recombination regions. On the other hand, a higher incidence of NAHR events can be expected in high recombination regions [3].

In this report, I explore whether inversions are evenly distributed among chromosomes and which variables affect the generation and/or maintenance of inversions as polymorphisms in certain regions of the genome.

2 Data

2.1 Inversions

We are using a curated dataset of inversions identified and validated with different methodologies.
We will be analyzing 133 inversions, 54 NAHR and 79 NH.

2.2 Repeats

From UCSC track “GenomicSuperDups”. These are segmental duplications (SDs) with size \geq 1kb and \geq 90% identity. The table contains the positions of both copies and information about the identity between them. For some analyses in this report, those SDs occurring within the same chromosome (intrachromosomal) were identified.

2.3 Chromosome sizes

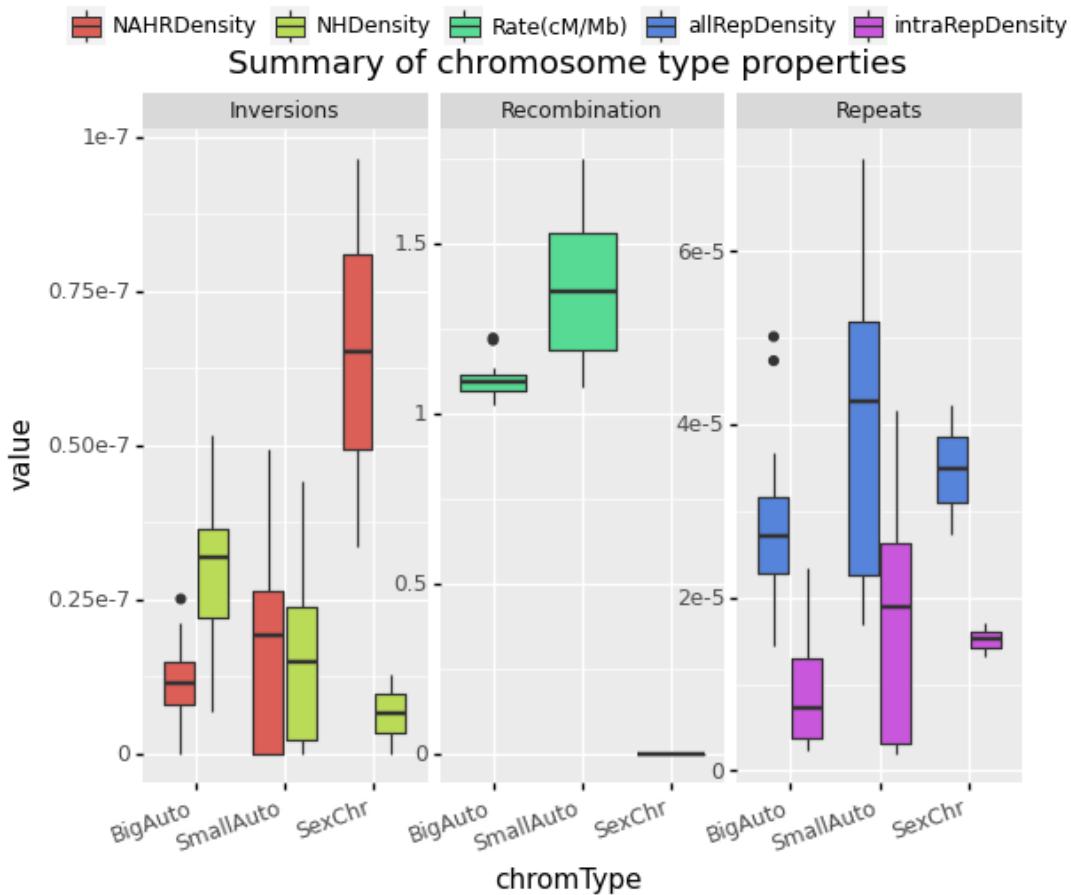
Total physical lengths for each chromosome are those specified for hg19 assembly in <https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37>.

Genetic lengths in centiMorgans were obtained from sex-averaged pedigree-based recombination maps [4] because they are easy to download and use, and have enough resolution for these analyses.

3 Analysis of whole autosomes

As suggested in previous meetings, for this analysis chromosomes were divided in two groups, big autosomes (1 to 12) and small autosomes (12 to 22). The reason behind this was to observe whether the differences between chromosome groups caused a difference on incidence of inversion types as well. As it can be observed in Figure 1, the only significant difference between chromosome types is that small autosomes have higher crossover rates, to make sure that at least two chiasmata are formed during meiosis. Within chromosome groups, big autosomes have significantly more NH than NAHR inversions, while small autosomes have equivalent proportions of both.

When looking at correlations between variables at the whole-chromosome scale (Figure 2), some patterns can be observed but given that there is much heterogeneity within chromosomes and that differences between chromosome types are so small, the analysis is noisy and lacks power.



<ggplot: (8774829223045)>

Figure 1: Summary of differences between chromosomes. Small autosomes are chr1 to 12 and big autosomes chr13 to 22. SexChr are X and Y.

R[write to console]: corrplot 0.92 loaded

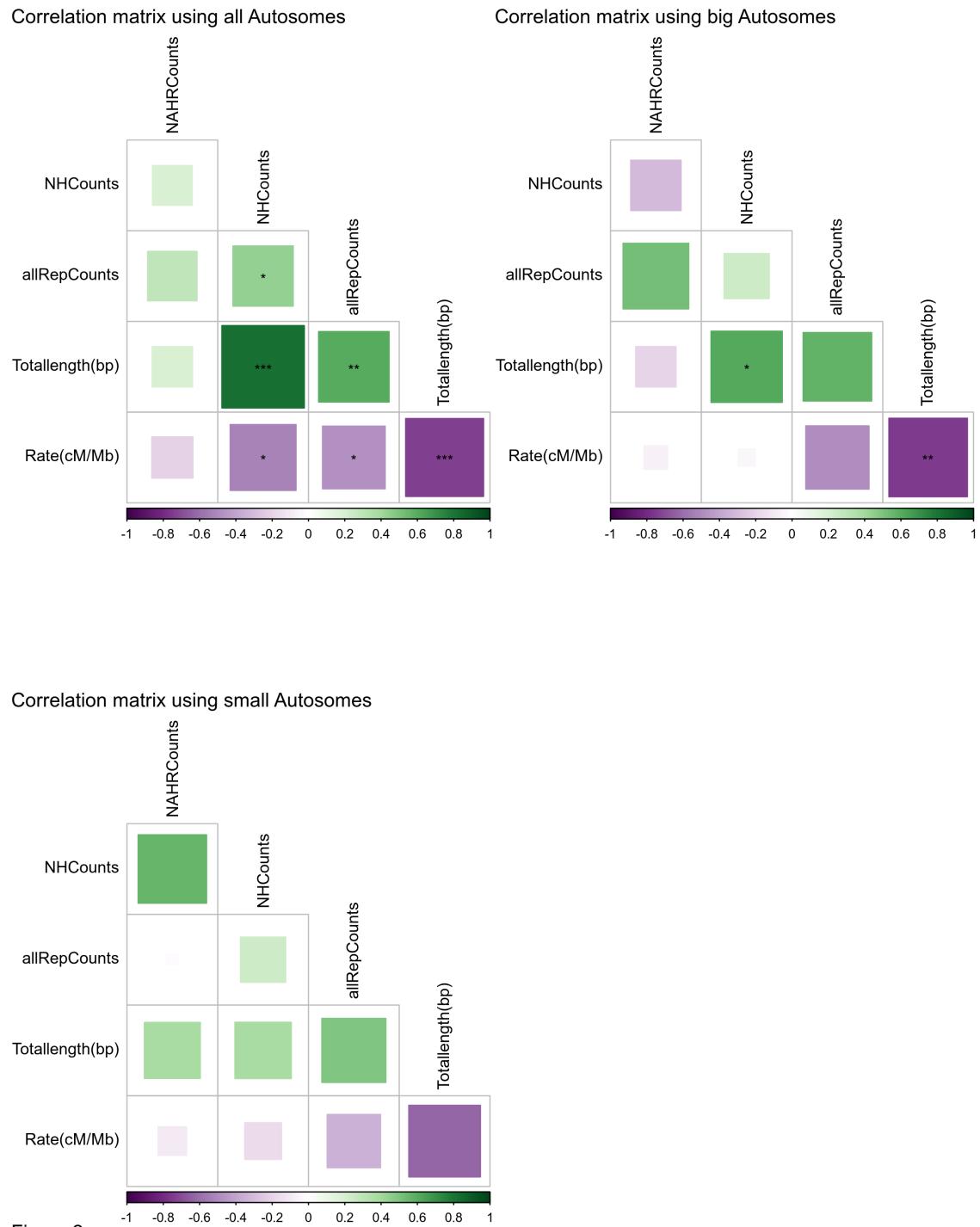


Figure 2

Big autosomes are chr1 to 12, small are chr13 to 22. The variables include Counts for inversions grouped by mechanism (NH, NAHR) and intrachromosomal repeats (intraRep), as well as chromosome physical and genetic length. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels.

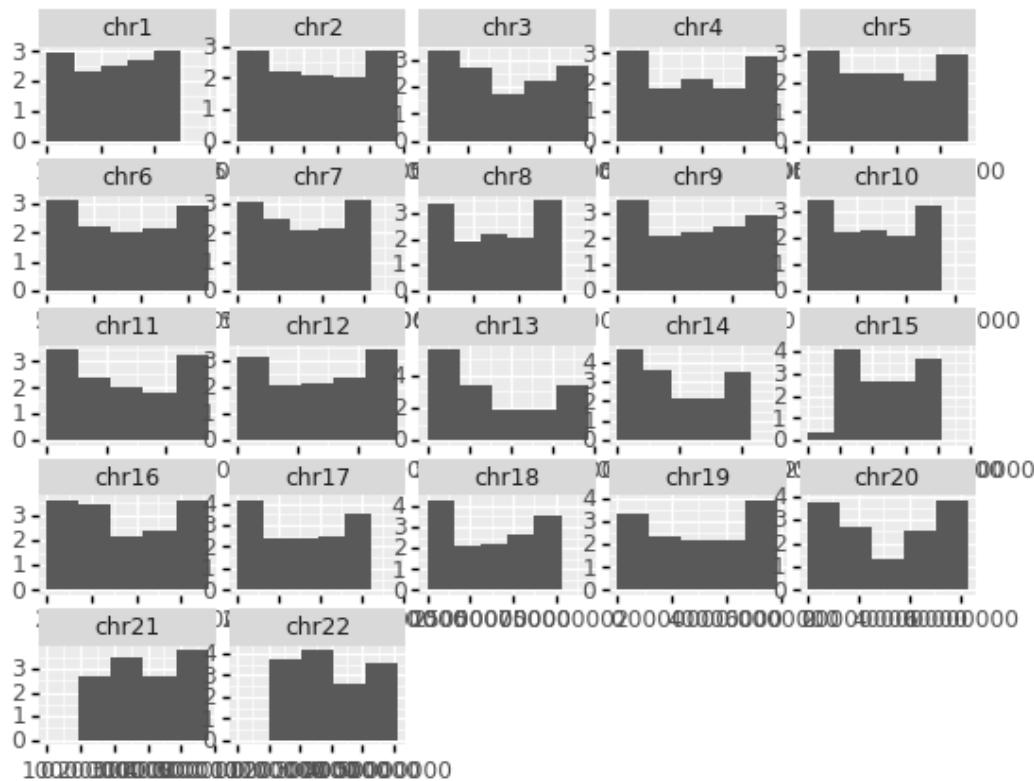
4 Analysis of autosomes divided in sections

Chromosomes were divided into 5 sections, which is enough to observe high-recombination regions (Figure 3) while keeping a range of windows with 0 to 5 inversions of either type. Sadly, they are not small enough to differentiate low recombination regions.

As shown in Figure 4, the quantity of NH inversions is directly proportional to window size and not influenced by the amount of repeats in the region, while the number of NAHR inversions in a window is proportional to how repetitive is the region, but not influenced by the window size, even when the quantity of repeats is. Supplementary Figure 11 shows that when looking at intrachromosomal repeats alone, the correlation is still significant. Larger windows are associated with lower recombination rates because window sizes are proportional to chromosome size and therefore keep the global tendency of having higher recombination rates in smaller chromosomes. Mean recombination rates of each window are not associated to a higher incidence of either inversion type at this scale.

These results suggest that the position of polymorphic inversions in the genome would be determined by their origin rather than the influence of surrounding recombination rates in their later failure or success. However, this may or may not be true at higher resolutions, which will be analyzed in the next report (Location Patterns 2).

Mean recombination rates for each chromosome region tested



<ggplot: (8774835593388)>

Figure 3: 5 bins per chromosome are small enough to differentiate high recombination regions (near telomeres), but not so clearly low recombination regions, which should be near centromeres.

Correlation matrix using autosomes divided into 5 windows

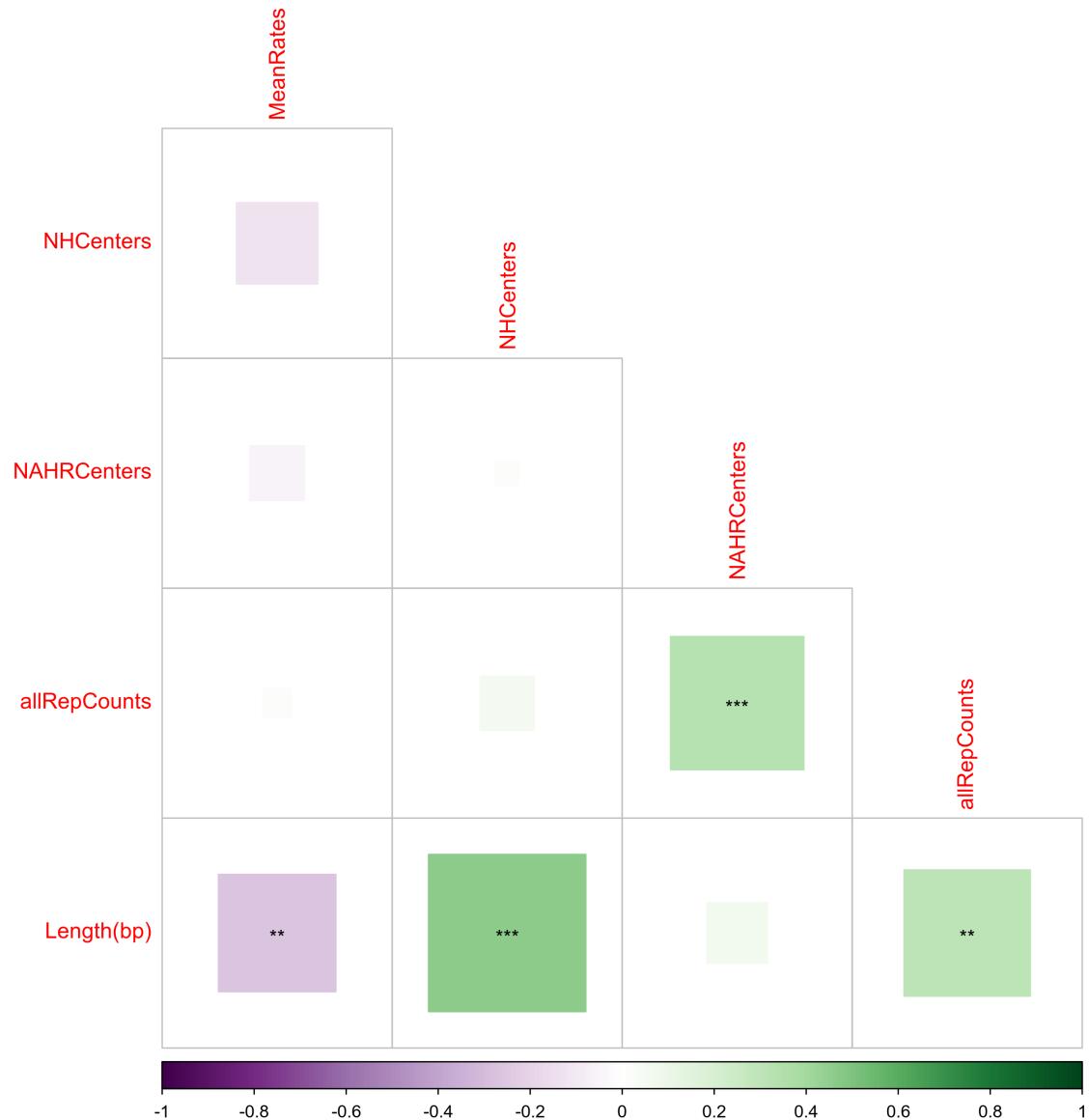
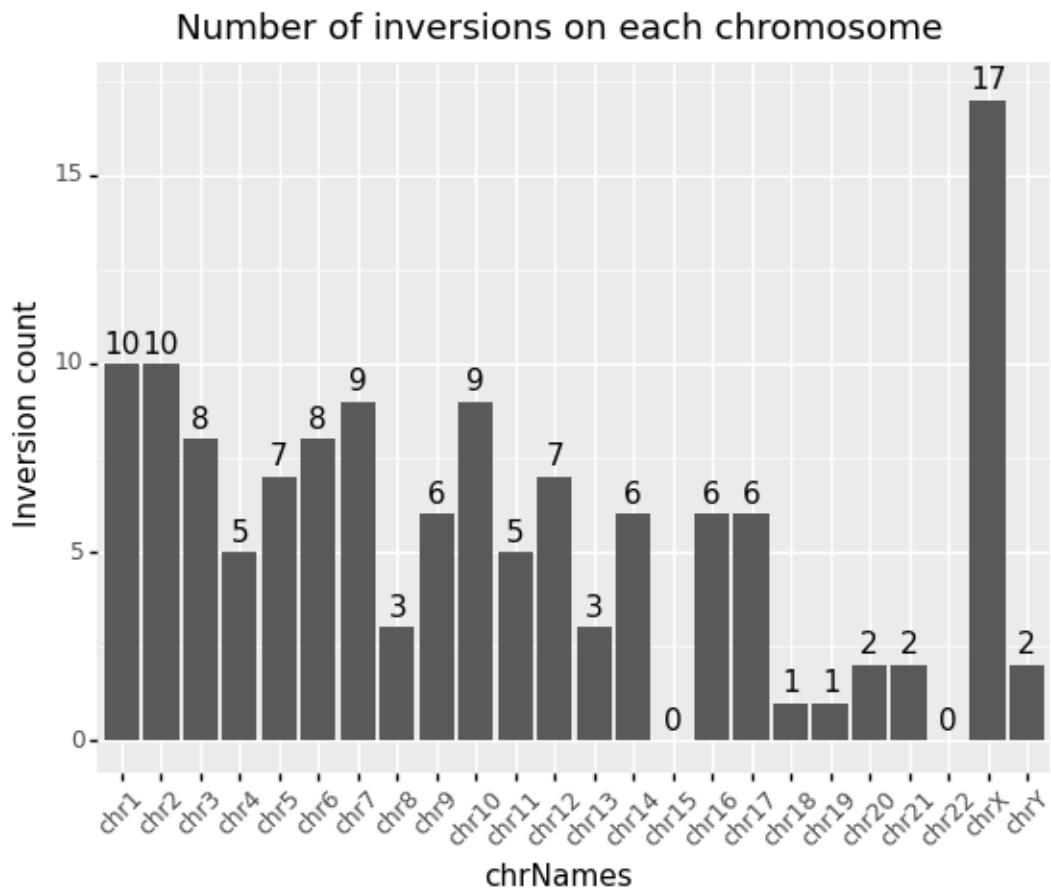


Figure 4

Each variable was calculated for autosomes divided into 5 windows. The variables include Counts of NH and NAHR inversion centers (Centers), Counts of all Repeats (allRep), window Length in bp and mean recombination rate (MeanRates). Green indicates positive correlation and purple negative correlation. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

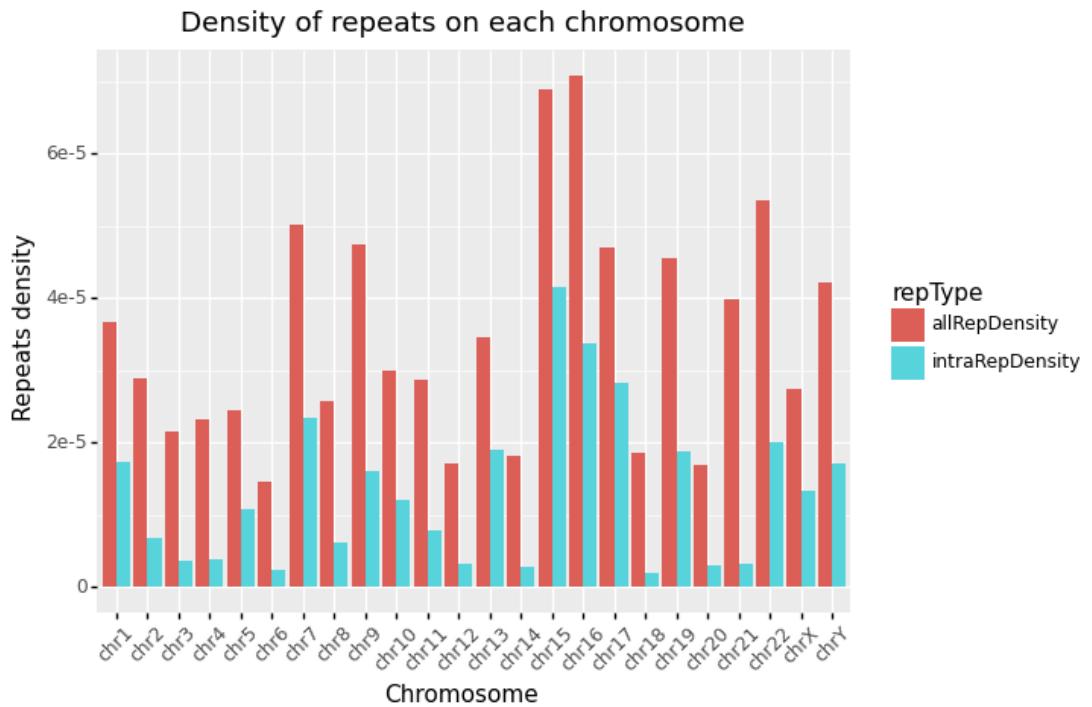
5 Supplementary figures

5.1 Differences between chromosomes in the distribution of variables



```
<ggplot: (8774829209793)>
```

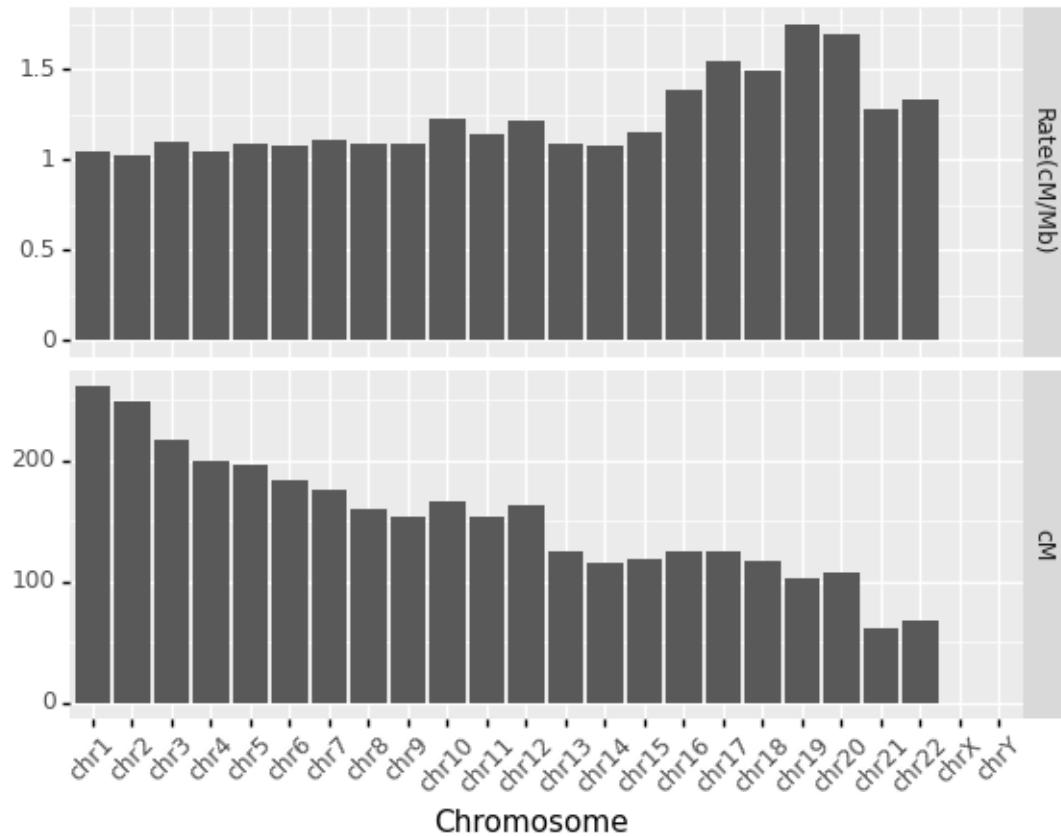
Figure 5: Number of inversions for each chromosome in the dataset.



<ggplot: (8774838995950)>

Figure 6: Desnity of repeats for each chromosome in the dataset.

Global recombination measurements for each chromosome



<ggplot: (8774829202122)>

Figure 7: Recombination rate and genetic length for each chromosome in the dataset.

5.2 Simple correlation matrices (just values)

Correlation matrix using all Autosomes



Figure 8

Each variable was calculated for each autosome. Counts are number of incidences, Density are Counts/Total chromosome length. The variables include Counts and Densities for Total inversions (inv), inversions grouped by generation mechanism (NH, NAHR), all Repeats (allRep) and intrachromosomal repeats (intraRep), as well as total physical and genetic chromosome length. Green indicates positive correlation and purple negative correlation. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Correlation matrix using Big Autosomes

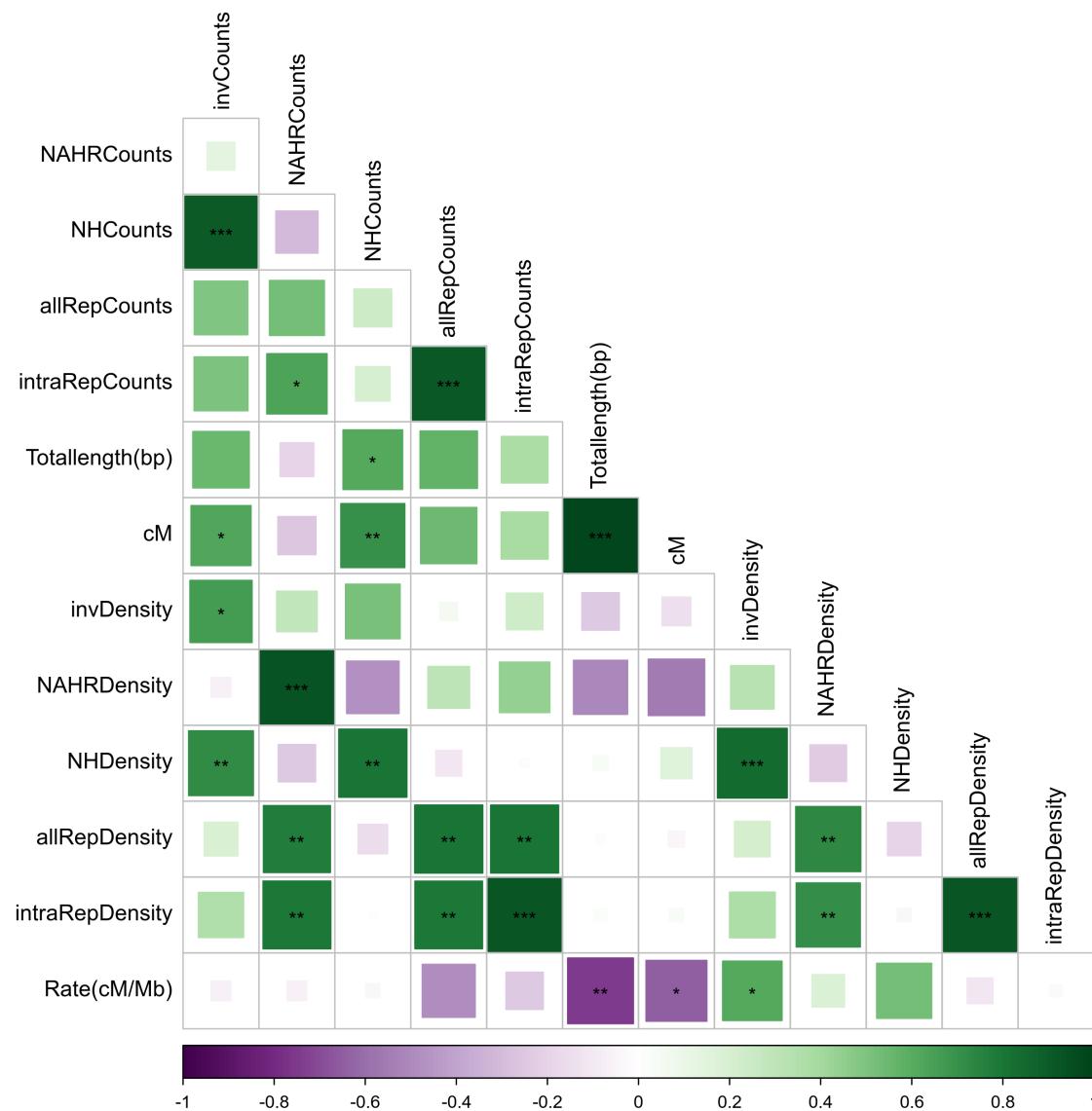


Figure 9

Each variable was calculated for autosomes 1 to 12. Counts are number of incidences, Density are Counts/Total chromosome length. The variables include Counts and Densities for Total inversions (inv), inversions grouped by generation mechanism (NH, NAHR), all Repeats (allRep) and intrachromosomal repeats (intraRep), as well as total physical and genetic chromosome length. Green indicates positive correlation and purple negative correlation. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Correlation matrix using Small Autosomes

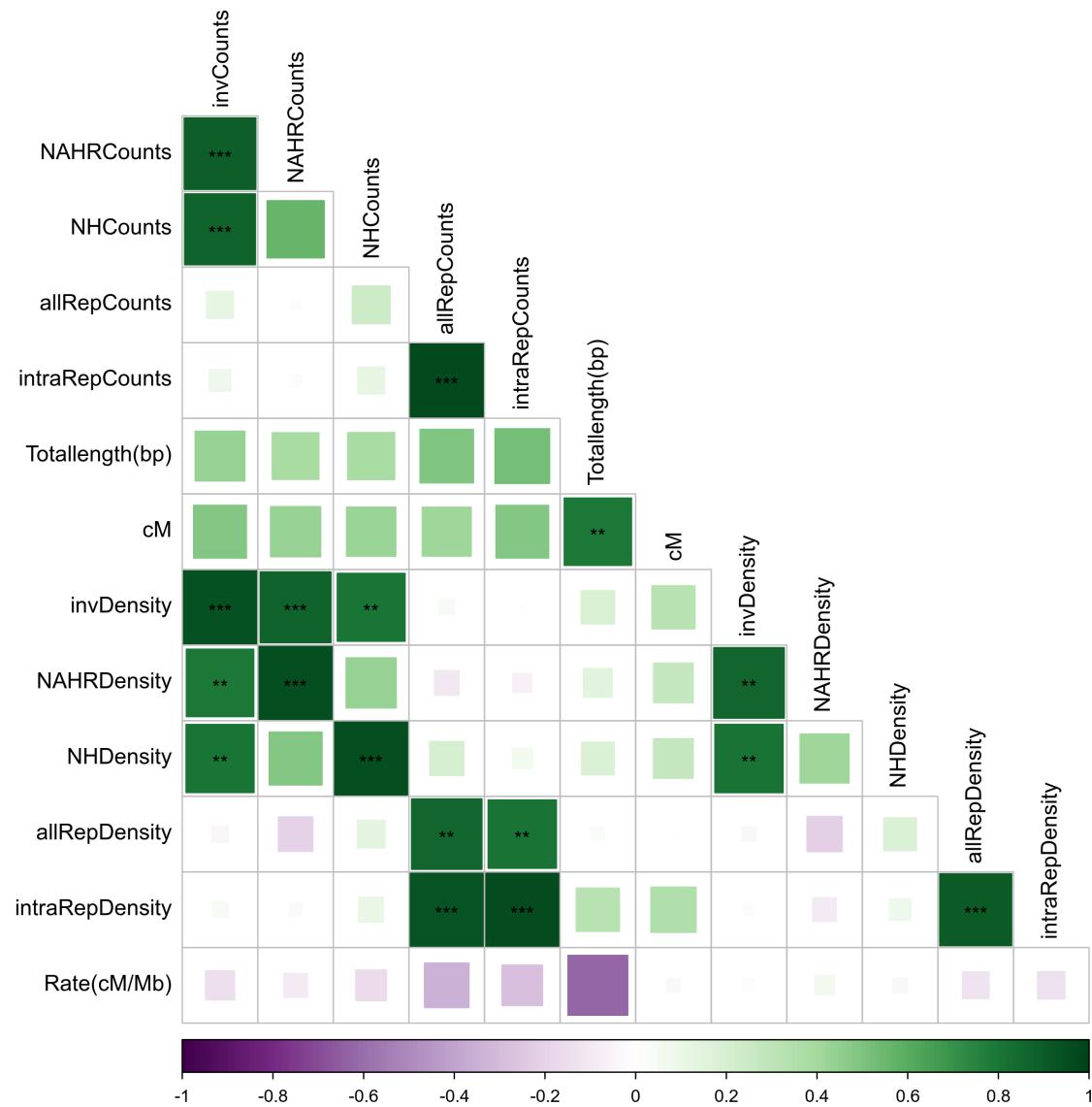


Figure 10

Each variable was calculated for autosomes 13 to 22. Counts are number of incidences, Density are Counts/Total chromosome length. The variables include Counts and Densities for Total inversions (inv), inversions grouped by generation mechanism (NH, NAHR), all Repeats (allRep) and intrachromosomal repeats (intraRep), as well as total physical and genetic chromosome length. Green indicates positive correlation and purple negative correlation. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Correlation matrix using autosomes divided into 5 windows

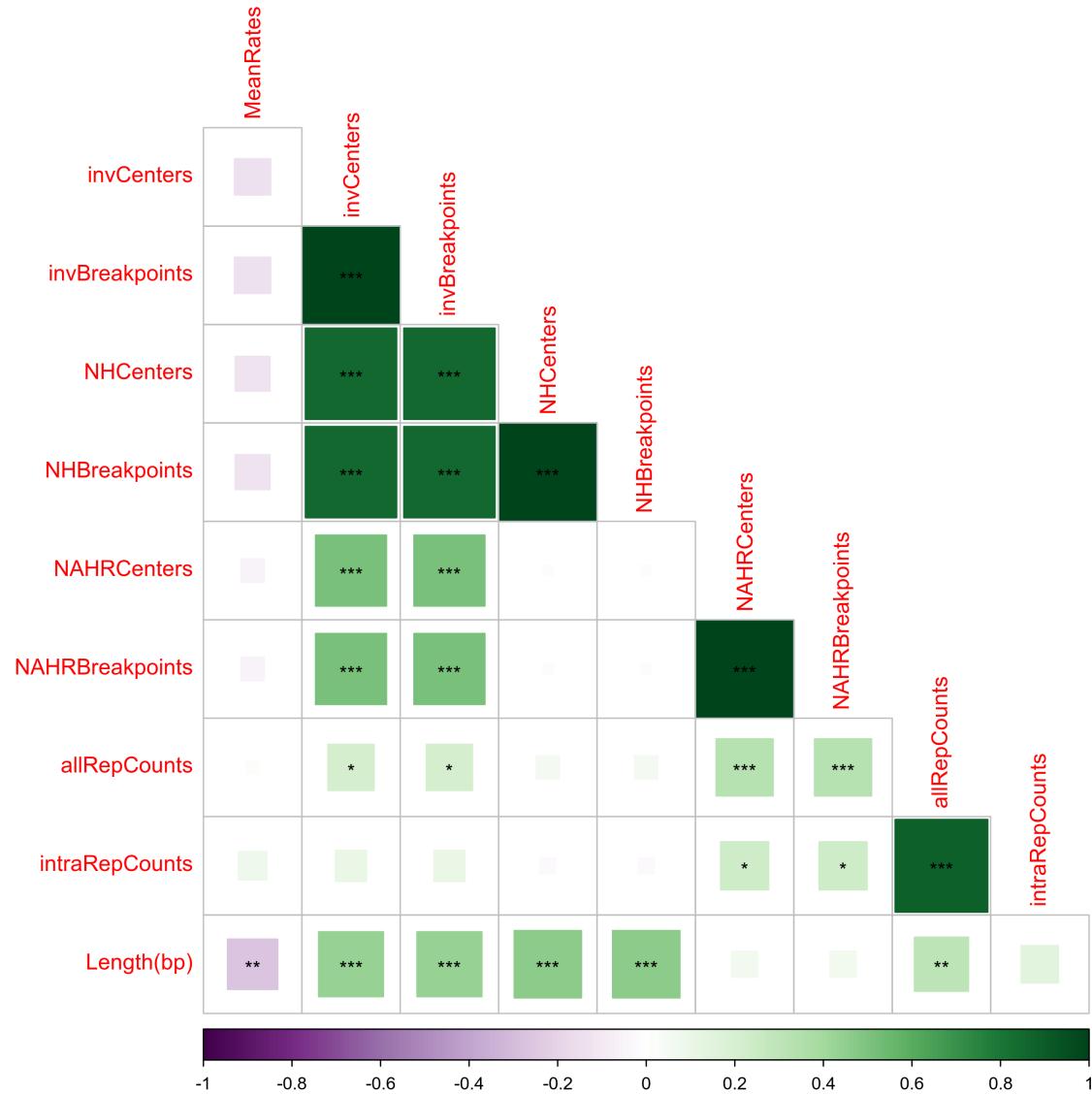


Figure 11

Each variable was calculated for autosomes divided into 5 windows. The variables include Counts of inversion centers (Centers) and breakpoint centers (Breakpoints) for all inversions (inv) and for inversions grouped by generation mechanism (NH, NAHR), Counts of all Repeats (allRep) and intrachromosomal repeats (intraRep), and window Length in bp and mean recombination rate (MeanRates). Densities not included because they didn't seem to be useful. Green indicates positive correlation and purple negative correlation. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

5.3 Comprehensive correlation matrices (with distributions and scatter plots)

```
R[write to console]: Loading required package: xts  
R[write to console]: Loading required package: zoo  
R[write to console]:  
Attaching package: 'zoo'  
  
R[write to console]: The following objects are masked from 'package:base':  
as.Date, as.Date.numeric  
  
R[write to console]:  
Attaching package: 'PerformanceAnalytics'  
  
R[write to console]: The following object is masked from 'package:graphics':  
legend
```

Complete correlation matrix using all Autosomes

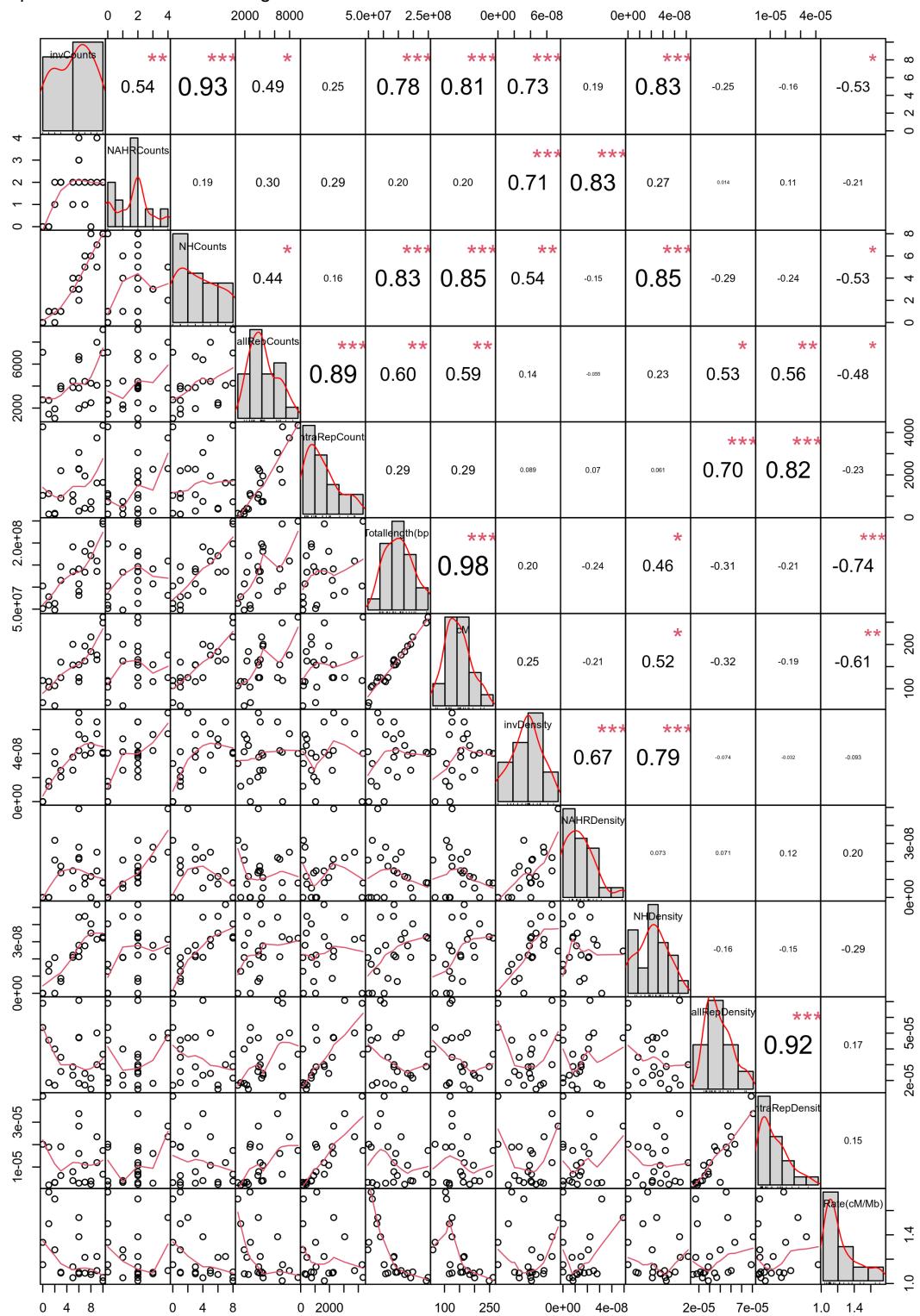


Figure 12:
Correlation matrix using all autosomes. Variable distributions on the diagonal, bivariate scatter plots with a fitted lines in the bottom half, and value of the correlation and significance level on the top half. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Complete correlation matrix using Big Autosomes

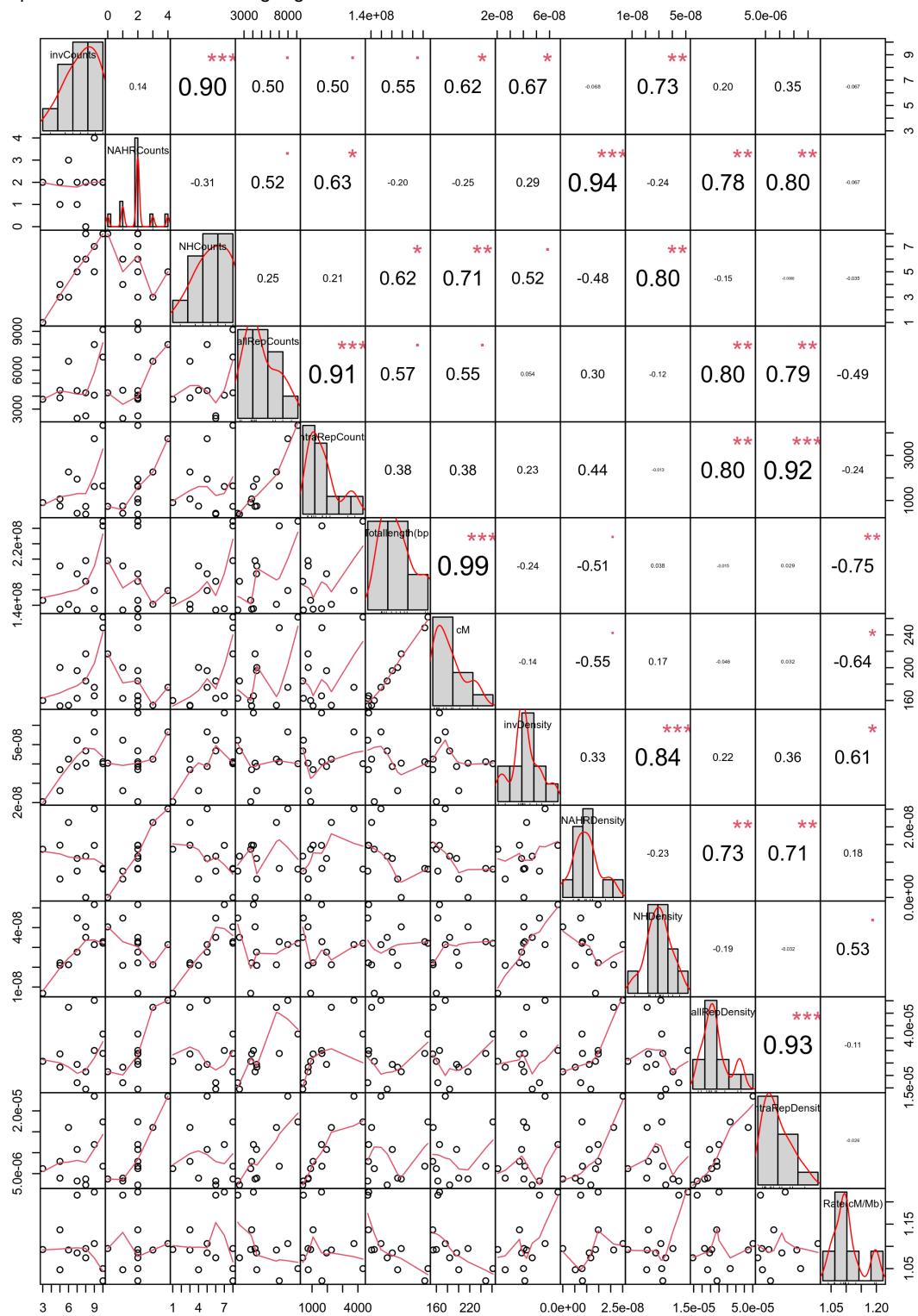


Figure 13:
Correlation matrix using autosomes 1 to 12. Variable distributions on the diagonal, bivariate scatter plots with a fitted lines in the bottom half, and value of the correlation and significance level on the top half. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Complete correlation matrix using Small Autosomes

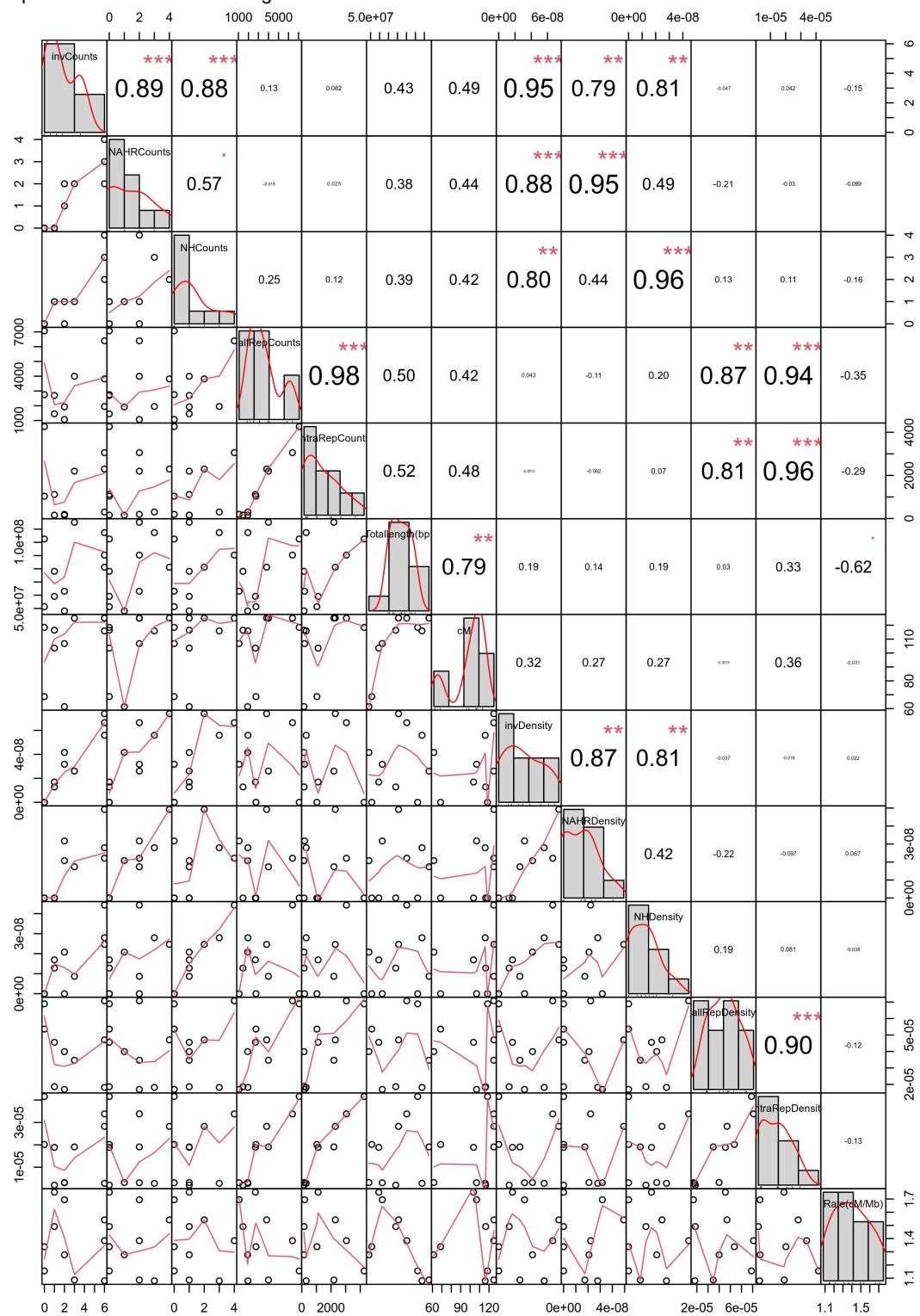


Figure 14:
Correlation matrix using autosomes 13 to 22. Variable distributions on the diagonal, bivariate scatter plots with a fitted lines in the bottom half, and value of the correlation and significance level on the top half. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

Complete correlation matrix using Autosomes divided into 5 windows

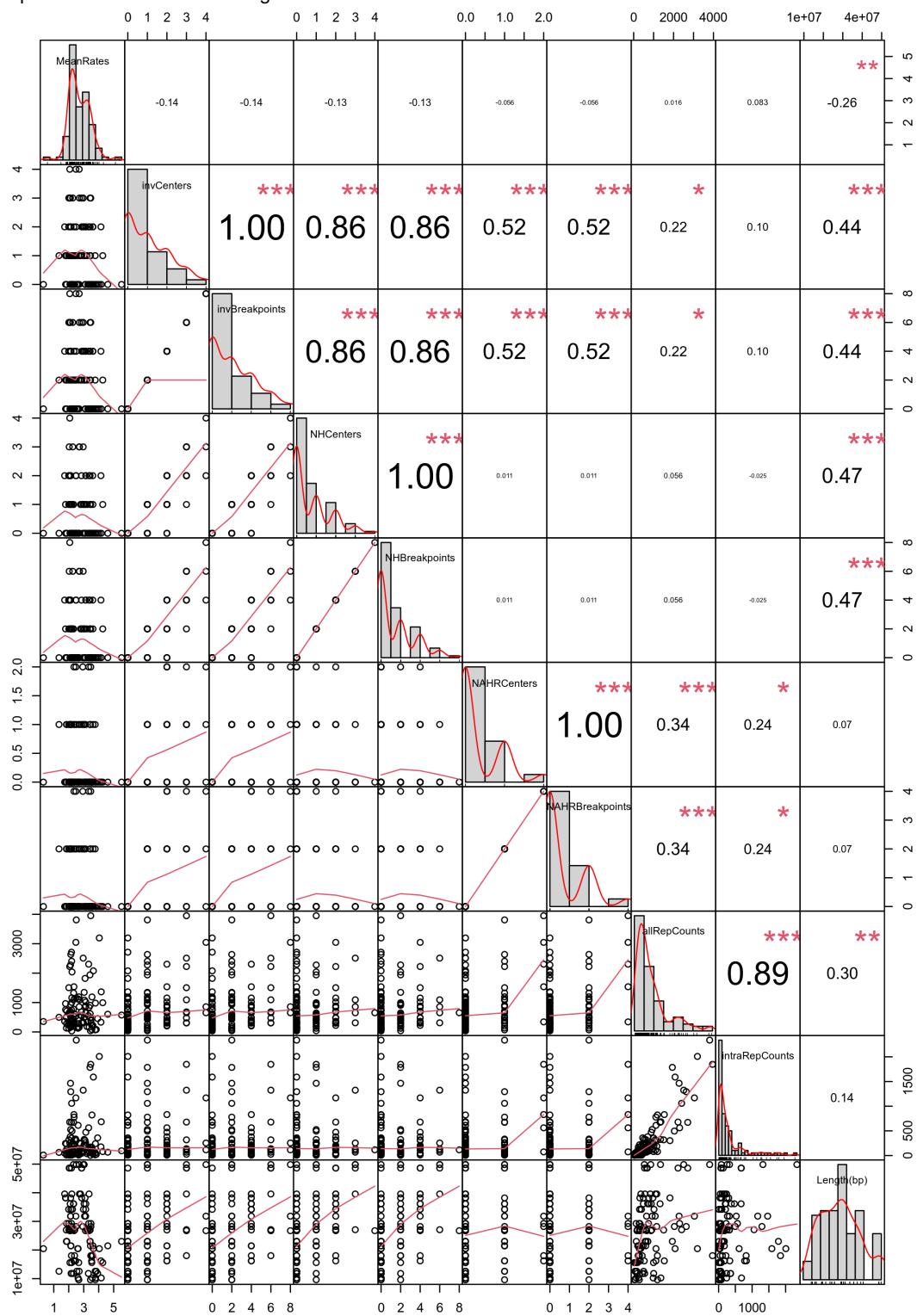


Figure 15:
 Correlation matrix using all autosomes, divided into 5 windows. Variable distributions on the diagonal, bivariate scatter plots with a fitted lines in the bottom half, and value of the correlation and significance level on the top half. One, two and three asterisks correspond to 0.05, 0.01 and 0.001 significance levels, respectively.

5.4 Previous test : not comprehensive correlations with chromosome characteristics

R[write to console]: Loading required package: ggplot2

R[write to console]: `geom_smooth()` using formula 'y ~ x'

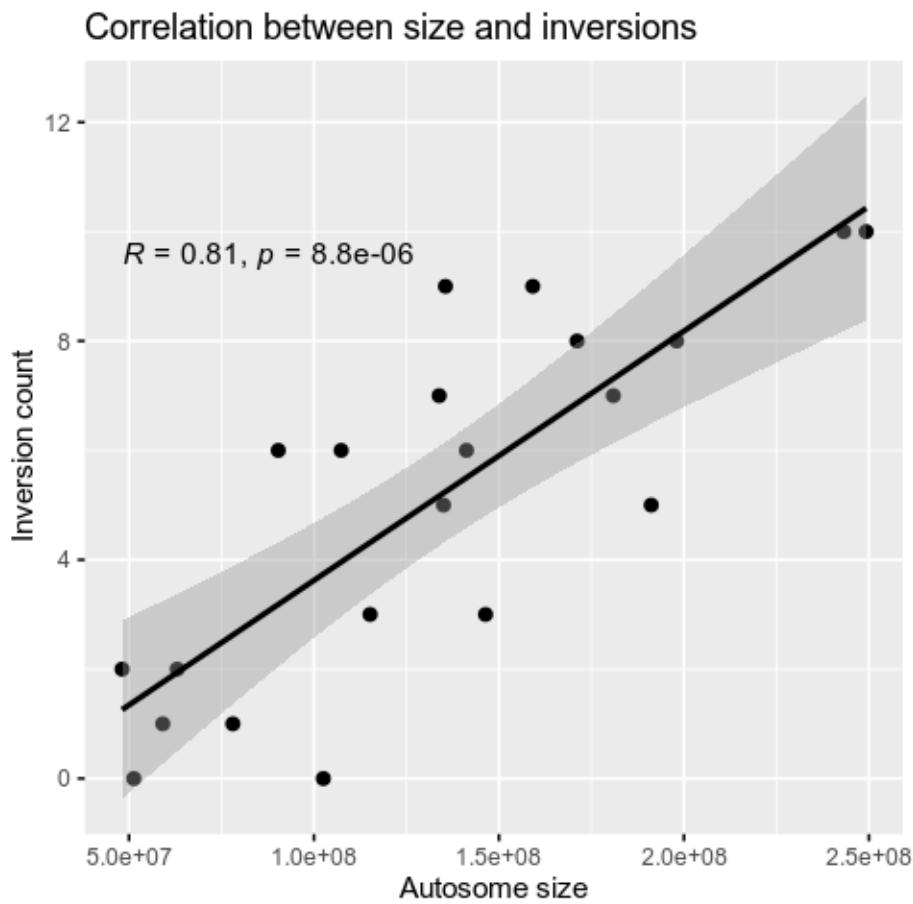


Figure 16: There is a moderate, significant positive correlation between autosome size and inversion count when all inversions and autosomes are considered.

```
R[write to console]: `geom_smooth()` using formula 'y ~ x'
```

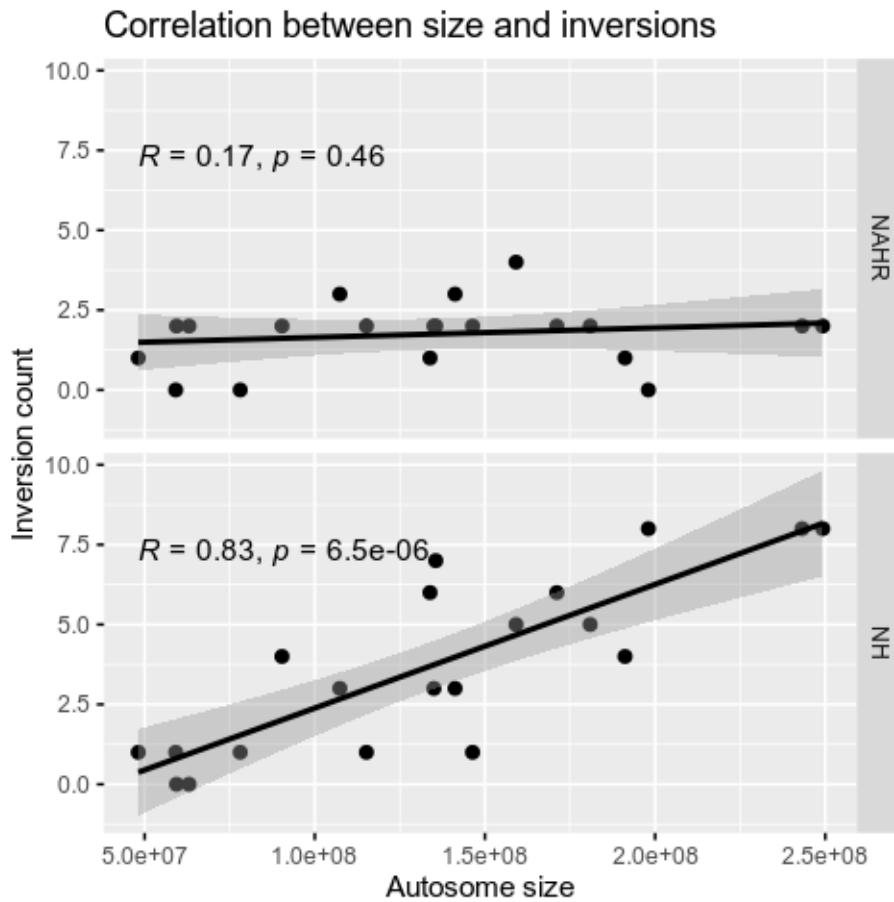
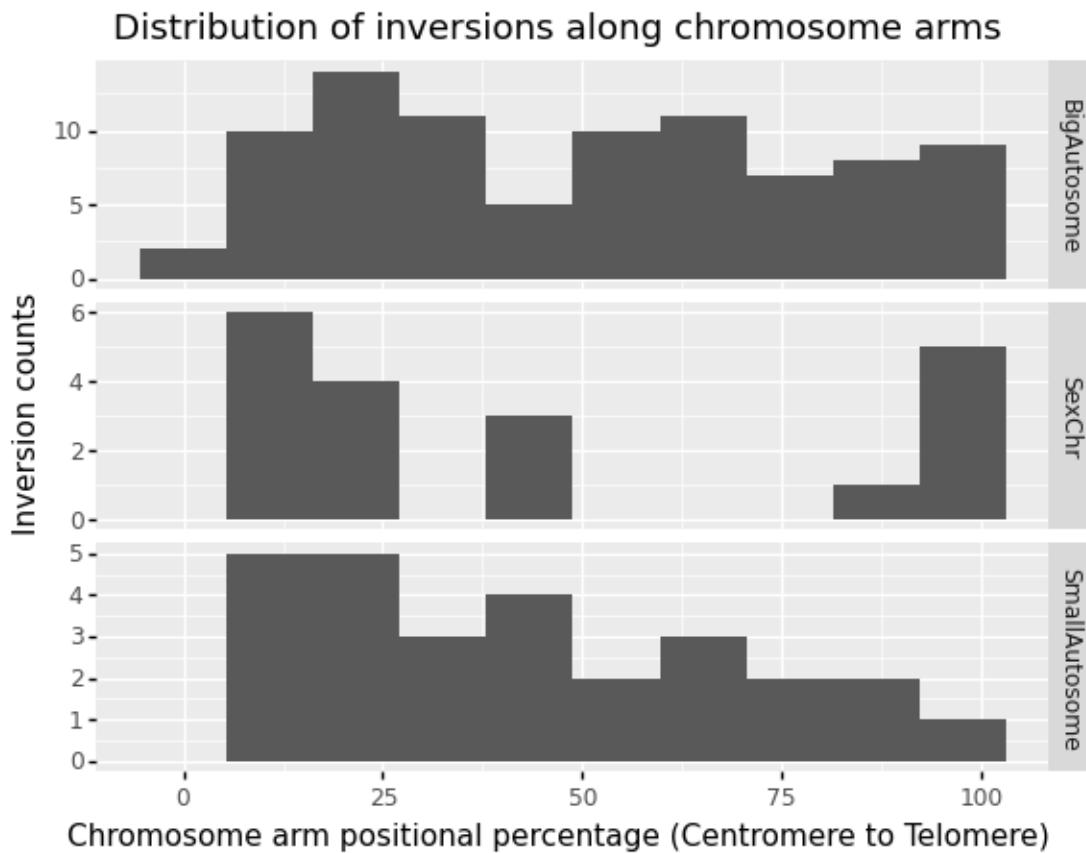


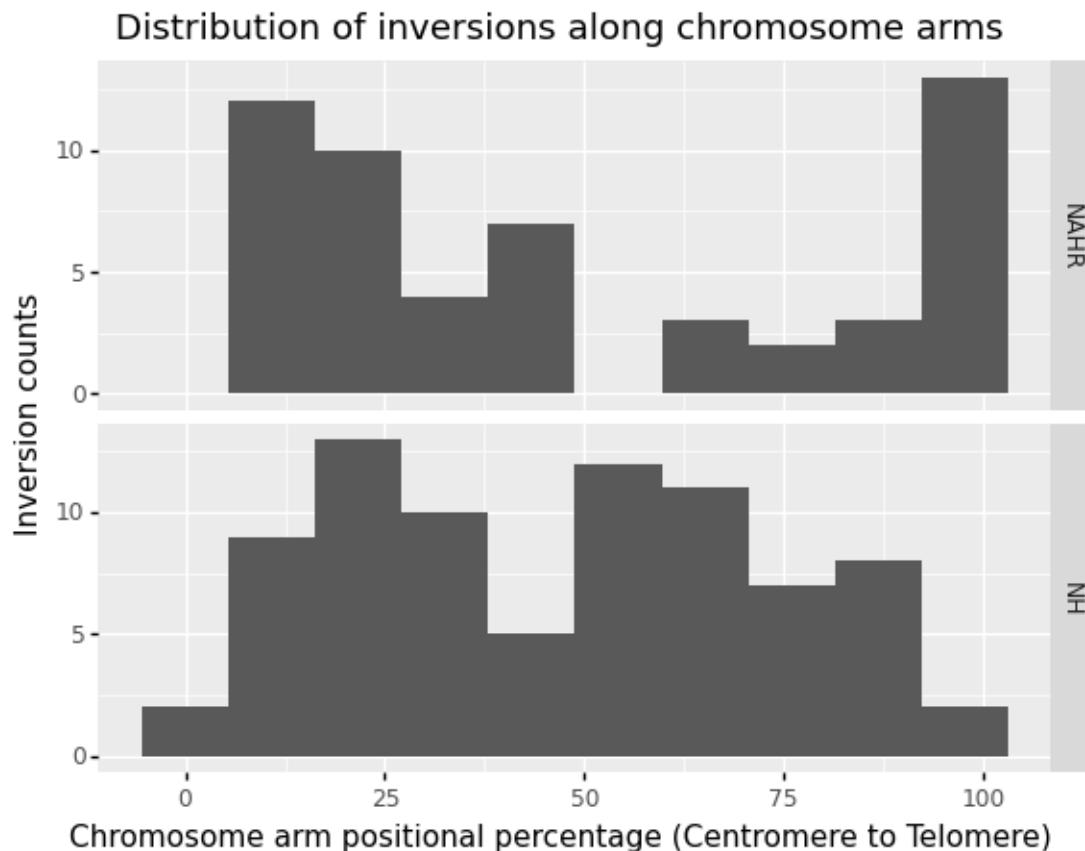
Figure 17: The correlation between Autosome size and inversion count is affected by the differential distributions observed in Figure 16. NH inversion counts show a strong, significant positive correlation with chromosome size, while no such correlation is observed for NAHR inversions because of their excess in Small chromosomes and deplete Big chromosomes.

5.5 Previous test: Distribution along chromosome arms



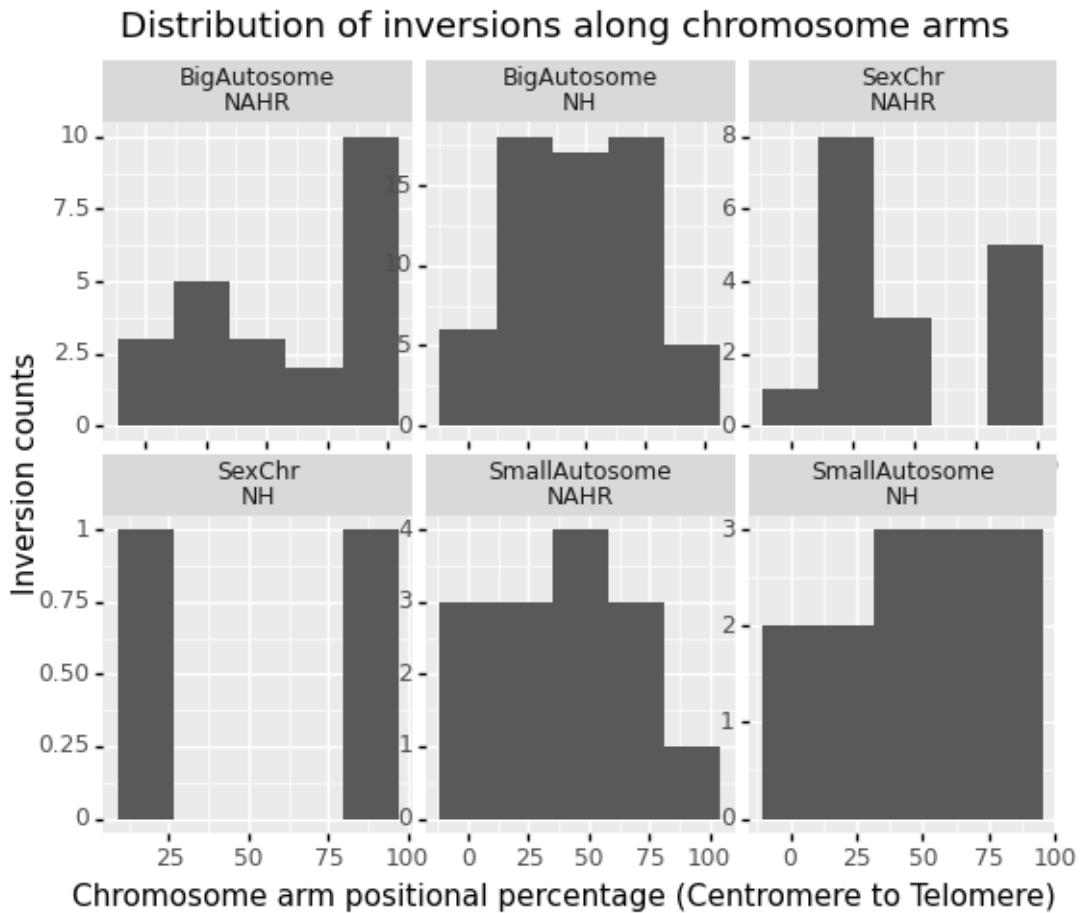
```
<ggplot: (8786104154188)>
```

Figure 18: Distribution of inversion counts along chromosome arms, where 0 is the centromere and 100 is the telomere. Inversions seem to be located preferentially around centromeres when not taking into account inversion origin, although in Sex chromosomes and Small Autosomes the first bin, most proximal to the centromere, is empty.



<ggplot: (8786104154197)>

Figure 19: Distribution of inversion counts along chromosome arms, where 0 is the centromere and 100 is the telomere. NAHR inversions are preferentially located near centromeres and telomeres, while NH inversions seem to be more dsitributed along the chromosome arm.



<ggplot: (8786104163892)>

Figure 20: Distribution of inversion counts along chromosome arms, where 0 is the centromere and 100 is the telomere. When looking at all the categories, we loose definition, especially for small chromosomes. The most evident patterns can be observed in Big chromosomes: NH inversions concentrate in the middle of the chromosome arm while most NAHR generate near telomeres.

References

- [1] Cristina Aguado, Magdalena Gayà-Vidal, Sergi Villatoro, Meritxell Oliva, David Izquierdo, Carla Giner-Delgado, Víctor Montalvo, Judit García-González, Alexander Martínez-Fundichely, Laia Capilla, Aurora Ruiz-Herrera, Xavier Estivill, Marta Puig, and Mario Cáceres. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS genetics*, 10(3):e1004208, mar 2014.

- [2] Carla Giner-Delgado, Sergi Villatoro, Jon Lerga-Jaso, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, Lorena Pantano, Bárbara D. Bitarello, David Izquierdo, Isaac Noguera, Iñigo Olalde, Alejandra Delprat, Antoine Blancher, Carles Lalueza-Fox, Tõnu Esko, Paul F. O'Reilly, Aida M. Andrés, Luca Ferretti, Marta Puig, and Mario Cáceres. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, 10(1):4222, dec 2019.
- [3] Mariko Sasaki, Julian Lange, and Scott Keeney. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology*, 11(3):182–195, mar 2010.
- [4] Claude Bhérer, Christopher L Campbell, and Adam Auton. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, 8:14994, apr 2017.