Detection of large scale inversion location patterns with ordinal logistic regression

Ruth Gómez Graciani

Contents

Contents

1 Windows generation

First, we obtain the density distribution, and local minima and maxima for the CEU Spence recombination map (Figure 1).

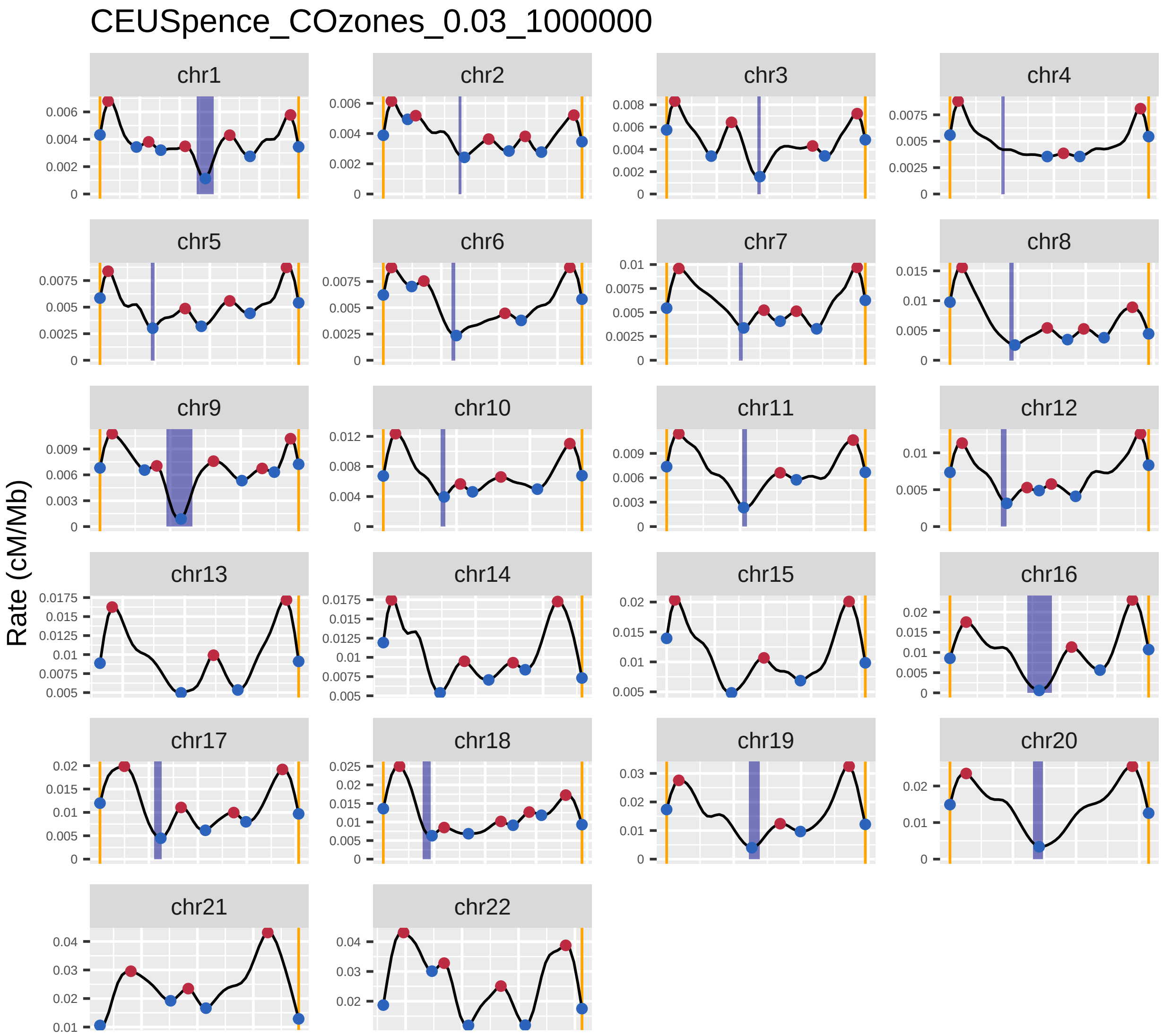

Figure 1: Black line is crossover density, blue and red points are local minima ans maxima respectively, centromeres in blue, chromosome limits in orange. Each chromosome has its own size and recombination scales. Telocentric chromosomes do not include p arm and centromere.

Next, we calculate telomeric and centromeric regions and divide the genome into windows accordingly (Figure 2). Telomeric regions are calculated from the extremes towards the centromere, starting at a chromosome limit and ending at the center point between the first local maximum and the next local minimum inwards. Centromeric regions are calculated from the centromere limits towards the extremes, starting at the centromere limit and ending at the center point between de limit and the next local maximum outwards. Centromeres are discarded because they tend to contain less reliable recombination estimates. The remaining spaces between telomeric and centromeric regions are divided into 2 windows and marked as arm regions. This last step is necessary, as arm regions are significantly larger than centromeric or telomeric regions if considered as a whole, which means that the levels of noise between the different chromosome regions would be different and could impact the result. In addition, this eliminates all the size outliers from the general distribution without actually deleting data (Figure 3).

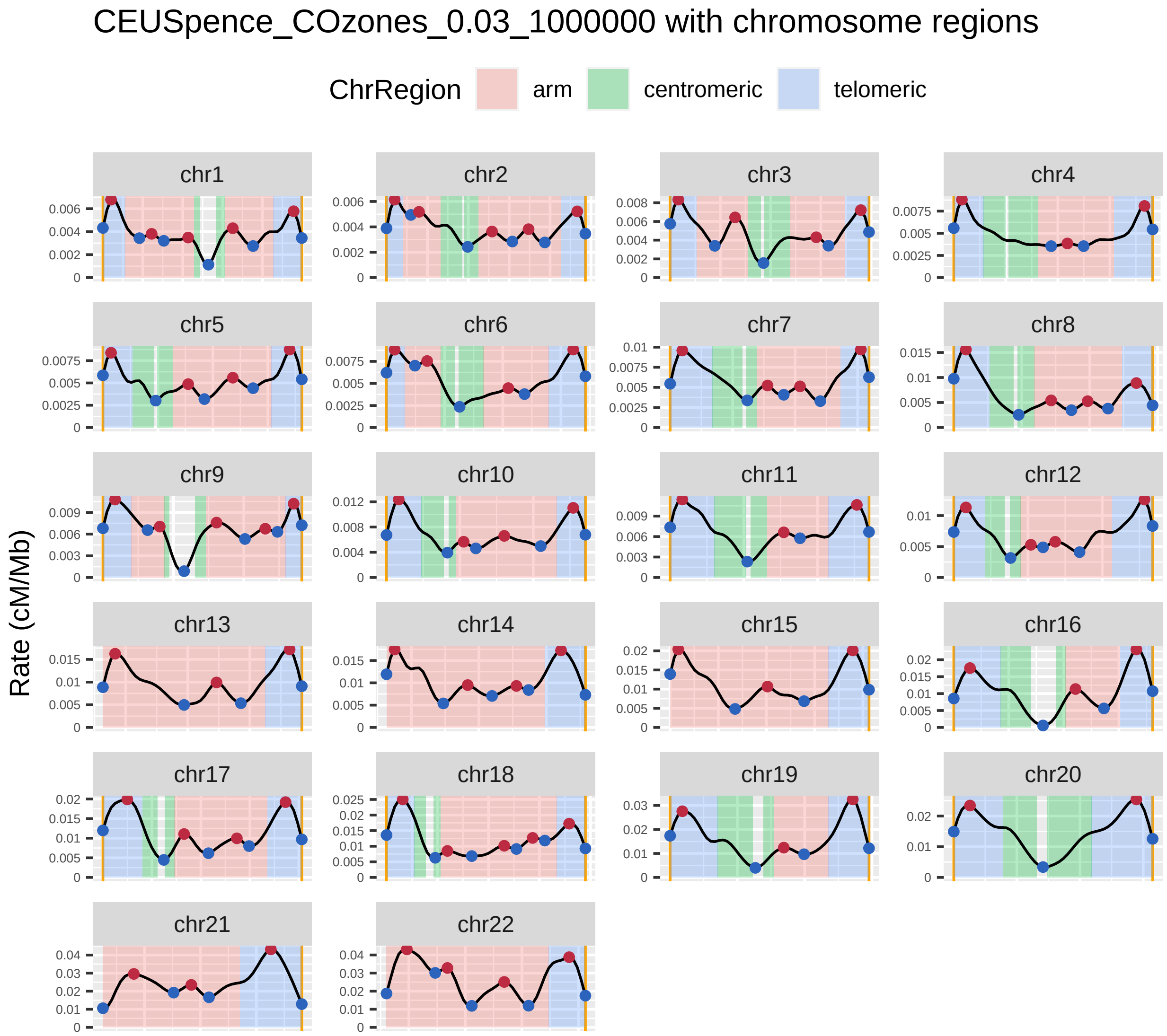

Figure 2: ChrRegion-coded windows for telomeric, centromeric and arm categories. Black line is crossover density, blue and red points are local minima ans maxima respectively, chromosome limits in orange. Each chromosome has its own size and recombination scales. Telocentric chromosomes do not include p arm and centromere.

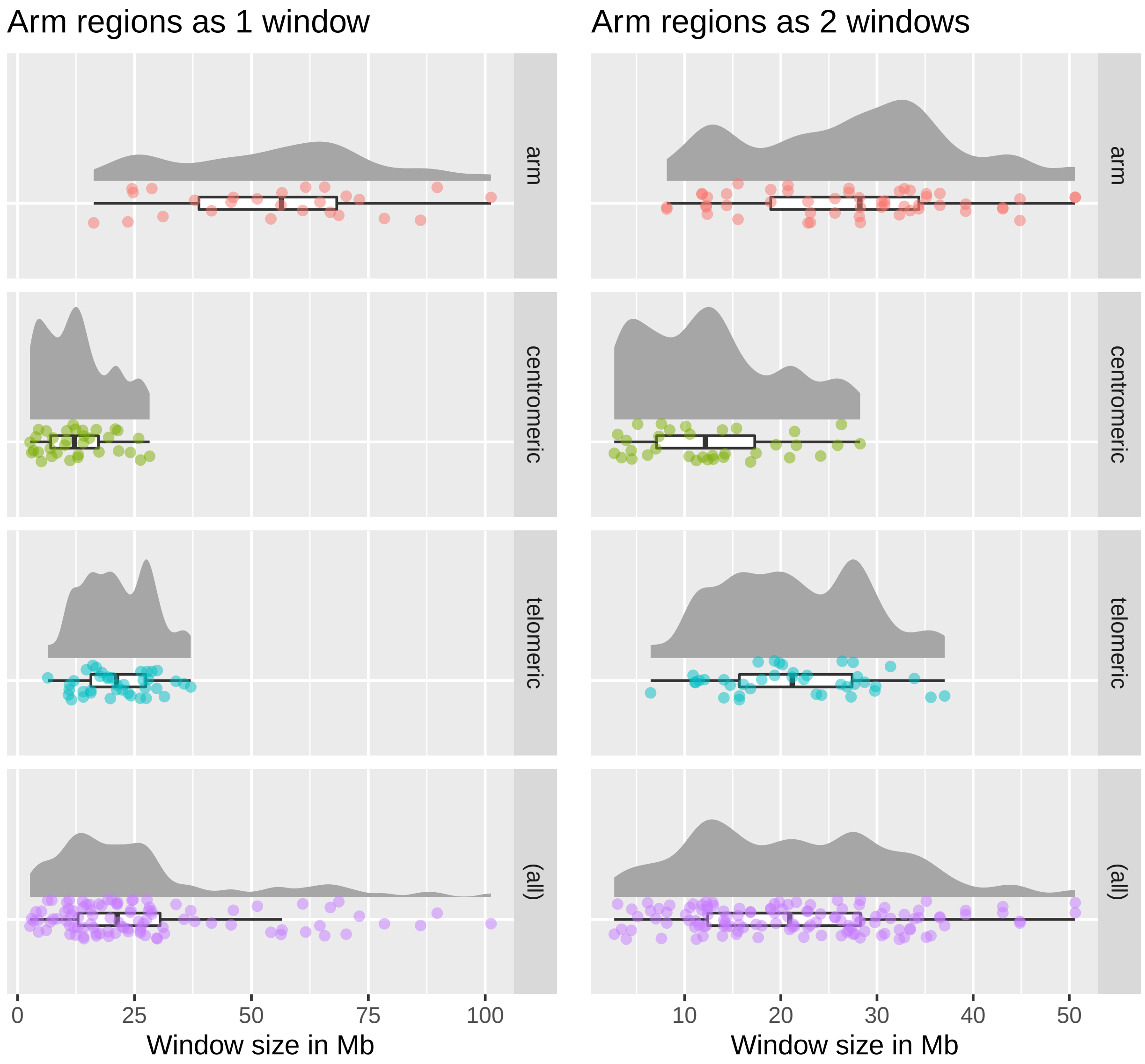

Figure 3: Raincloud plot for size distribution of windows depending on the treatment of arm regions. When arm regions are divided into 2 windows, the shape of the arm size distribution remains the same while the average is not significantly different from centromeric and telomeric regions. All outliers are removed from the general distribution without actually losing data.

2 Data preparation

2.1 Model assumptions

For each window, I calculated the number of total inversions, NH inversions, and NAHR inversions, the window length in Mb, number of repeats and the average recombination rate in cM/Mb. I want to perform Ordinal Logistic Regressions for all the inversions, NH, and NAHR inversions.

The assumptions of the Ordinal Logistic Regression are as follow:

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. No multi-collinearity.
4. Proportional odds.

I show the data distributions in the Figure 4. The inversion counts have only a number of possible options, so they can be considered an ordinal variable. Since there are only a few cases of some of the inversion count options (Table 1) I will make a “3 or more” category (count cases in Table 2). The independent variables are continuous and categorical, so assumptions 1 and 2 are satisfied.

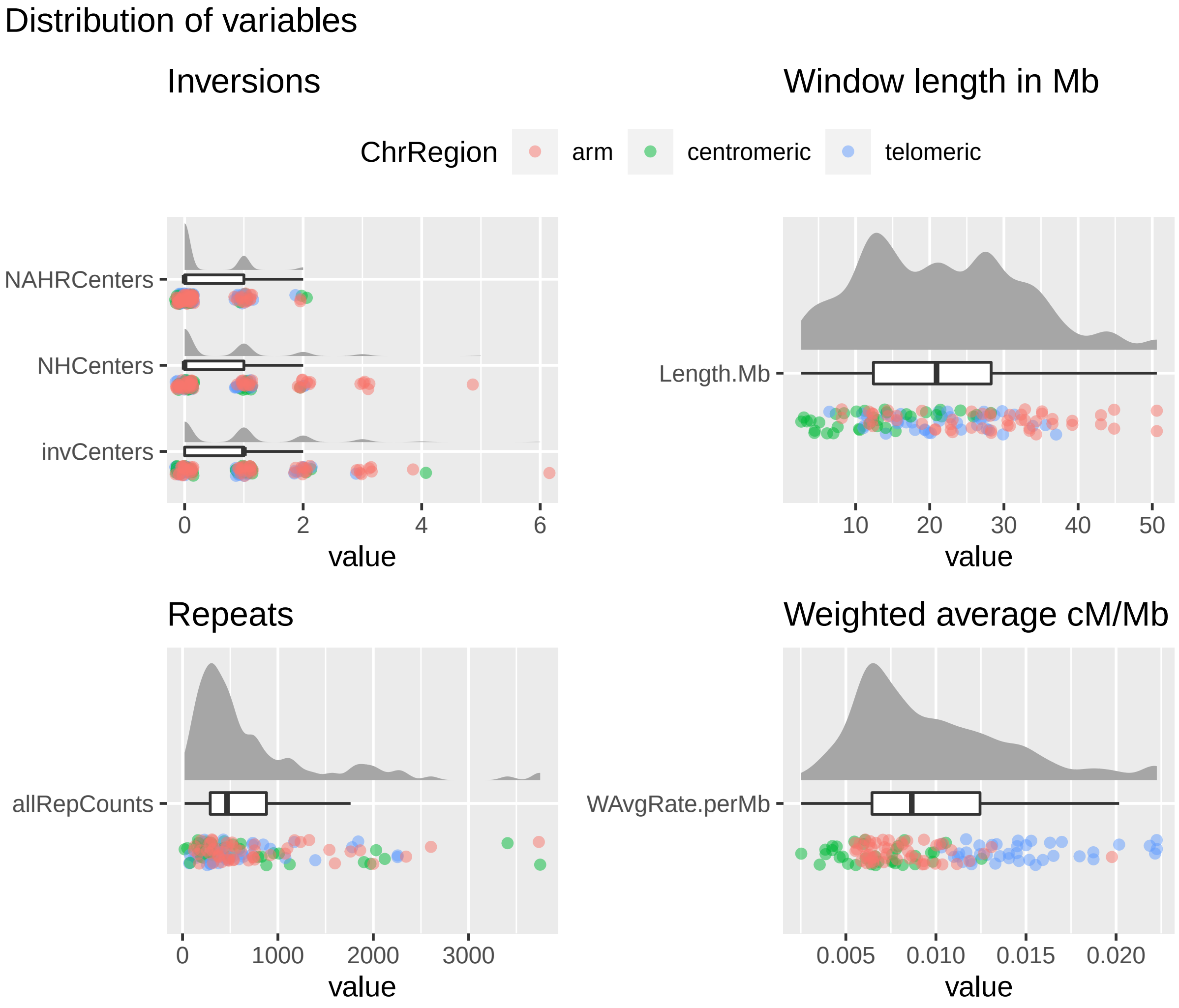

Figure 4: Raincloud plots for each variable.

Table 1: Original category counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	56	74	92
1	40	34	28
2	18	11	5
3	8	5	NA
4	2	NA	NA
5	NA	1	NA
6	1	NA	NA

Table 2: New category counts

CountGroups	invCategory	NHCategory	NAHRCategory
0	56	74	92
1	40	34	28
2	18	11	5
3+	11	6	NA

With these groups, I visualize the relationships between dependent and independent variables in Figure 5.

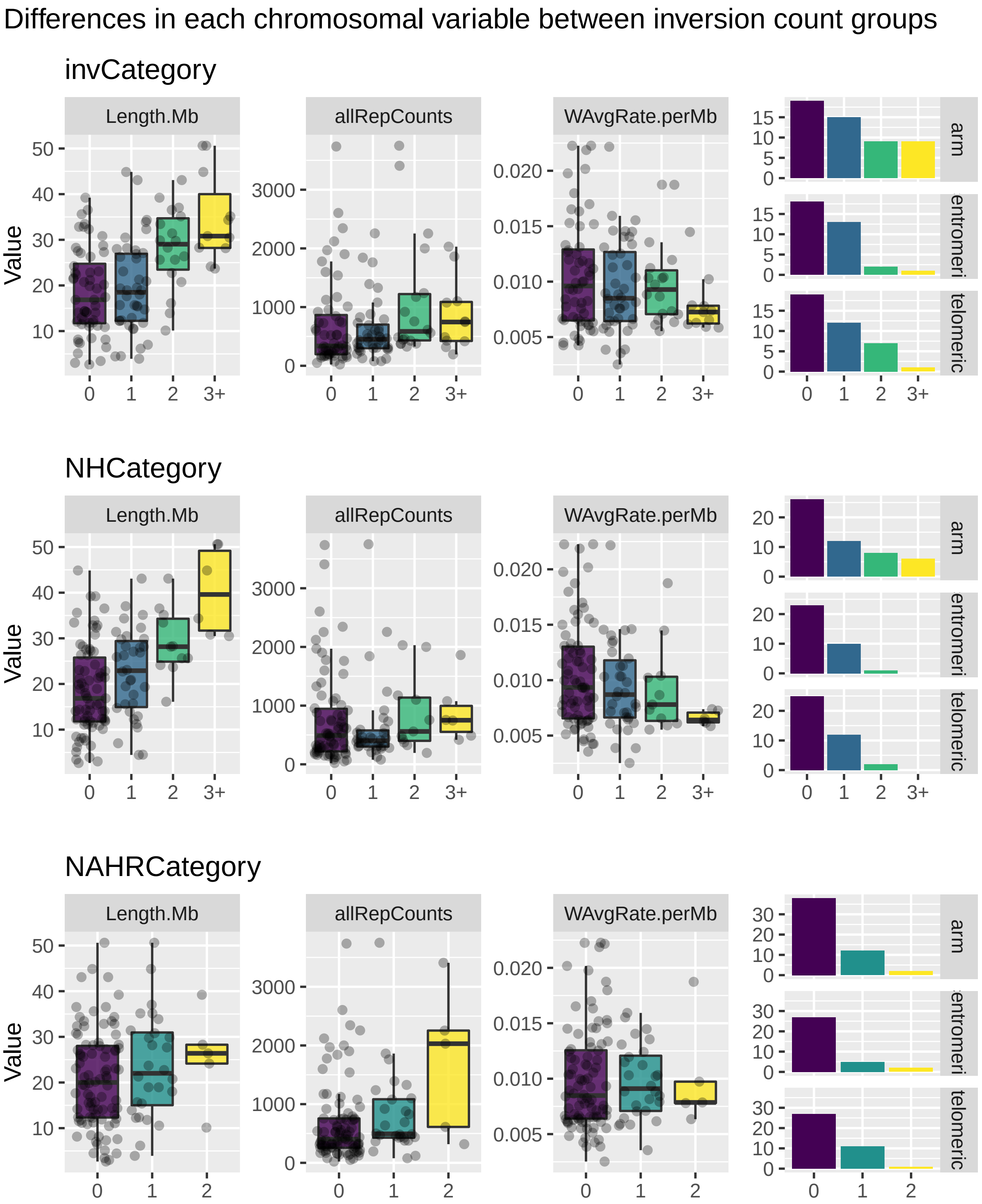

Figure 5: Boxplots for each dependent variable group and each independent variable quickly show candidates of having a strong effect. We can also see that there is missing data for some chromosome region types, because windows with 3+ inversions are scarce.

Finally, I will test assumption number 3, no multi-collinearity between independent variables. Figure 6 shows that some of the independent variables are significantly correlated, but this does not confirm multi-collinearity. I performed a variance inflation factor test on the correspongING linear model to further check the multi-collinearity (Table 3). The general rule of thumbs for VIF test is that if the VIF value is greater than 5, we should proceed with caution, and if the value is greater than 10, then there is multi-collinearity, so we can say that the third assumption (no multi-collinearity) is satisfied, but that we should be cautious when interpreting results involving the chromosome region variable. This result may be explained by the significantly higher recombination rate of telomere regions.

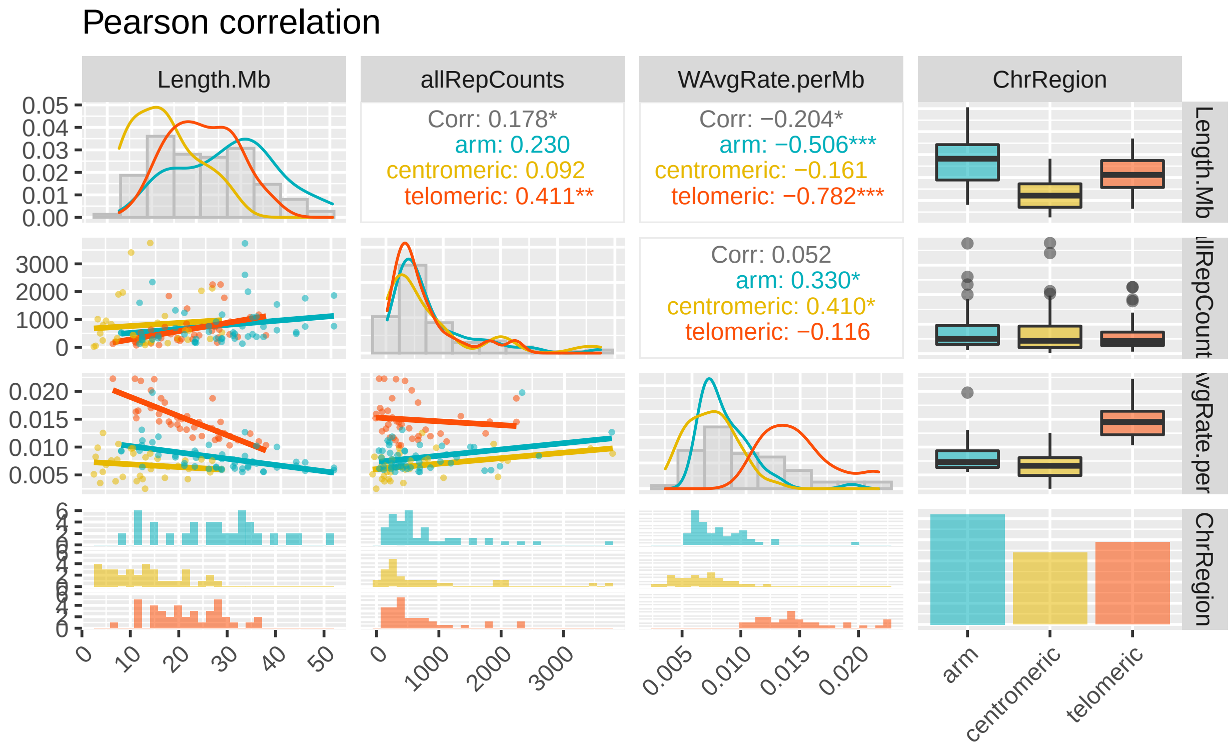

Figure 6: Correlation matrix of independent variables.

Table 3: Variance Inflation Factor

	GVIF	Df	GVIF^(1/(2*Df))
Length.Mb	2.272644	1	1.507529
allRepCounts	1.232203	1	1.110046
ChrRegion	5.610644	2	1.539052
WAvgRate.perMb	4.011360	1	2.002838

The proportional odds assumption will be tested for each model that we fit in the following analyses.

2.2 Scaling of distributions

Standardized coefficients are useful in multiple scenarios,for example, to compare effects of predictors reported in different units. In out case it is necessary because the `polr` function depends on methods that require data scalation for them to be the reliable. The most straightforward way is using the Agresti method of standardization, applied with the `scale()` function, which adjusts the mean to 0 and the standard deviation to 1. Once the model is fitted, we can use the standard deviation of the original distribution to transform scaled coefficients to natural coefficients and viceversa.

Table 4: Example of variable scalation.

	Length.Mb	Length.Mb.Scaled
Min.	2.680877	-1.7498964
1st Qu.	12.405960	-0.8498356
Median	20.918330	-0.0620119
Mean	21.588363	0.0000000
3rd Qu.	28.276348	0.6189760
Max.	50.611965	2.6861476

3 Model fitting

For all inversions, NH inversions and NAHR inversions, I have fitted an Ordinal Logistic Regression with `polr`, which returns a coefficient that represents the log(OddsRatio) and the corresponding Standard Error and t-value. For ease of interpretation, I have included in the table the Odds Ratio and the p-value corresponding to the t-value.

The summary tables also include the p-value for the Brant test, that checks whether the proportional odds assumption is true for this dataset (H0 = the proportional odds assumption holds). All models fulfill this assumption, although for the NAHR these p-values are to be taken with caution because some ChrRegion-invCategory combinations did not exist.

The interpretation for the Odds Ratio is better understood with an example: given an independent variable x with an Odds Ratio of 1.25, for each increase in 1 measurement unit of variable x, a window is 1.25 times more likely to be in a higher inversion count category, given that the other variables remain constant. Since our variables are scaled, measurement units are standard deviations of each variable distribution. A significant p-value means that within the coefficient's confidence interval the values do not change sign, i.e. the variable is consistently increasing or decreasing the odds.

In the general model (Table 5) the variable Length is the only significant one, with an Odds Ratio = 2.197 and p-value = 0.002. When only NH inversions are taken into account (Table 6), both the significance and the Odds Ratio increase: OR = 2.955 and p-value = 0.0002. On the other hand, when only NAHR inversions are considered (Table 7), the Length effect is no longer relevant and the only significant variable is Repeat number, with OR = 1.74 and p-value = 0.01. This confirms previous findings that at large window scales, the number of inversions depends on the generation mechanism, so the amount of repeats will be a key determinant of the amount of NAHR inversions in a region, while NH inversions will be generated randomly anywhere in the genome. It also evidences that when studying all inversions together, the patterns will be a mix of both inversion types, thus the importance of studying them separately given their very distinct behaviors.

Table 5: Model summary for invCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
allRepCounts.Scaled	0.1494831	1.1612339	0.2012853	0.7426430	0.4576978	0.2006168
ChrRegioncentromeric	-0.0769972	0.9258924	0.5877560	-0.1310020	0.8957737	0.8454186
ChrRegiontelomeric	-0.1400951	0.8692756	0.6149695	-0.2278081	0.8197954	0.3968334
Length.Mb.Scaled	0.7872092	2.1972558	0.2621008	3.0034602	0.0026693	0.0987999
WAvgRate.perMb.Scaled	-0.1573322	0.8544202	0.3544810	-0.4438381	0.6571596	0.6034830

Table 6: Model summary for NHCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
allRepCounts.Scaled	-0.1908700	0.8262400	0.2186493	-0.8729505	0.3826900	0.9902032
ChrRegioncentromeric	0.2980609	1.3472439	0.6486845	0.4594852	0.6458858	0.8283372
ChrRegiontelomeric	-0.5733191	0.5636515	0.6701624	-0.8554929	0.3922783	0.1562021
Length.Mb.Scaled	1.0836235	2.9553689	0.2975086	3.6423271	0.0002702	0.3627909
WAvgRate.perMb.Scaled	0.0632156	1.0652564	0.4057628	0.1557944	0.8761951	0.2698118

Table 7: Model summary for NAHRCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
allRepCounts.Scaled	0.5582261	1.7475697	0.2224411	2.5095450	0.0120887	0.9181911
ChrRegioncentromeric	-0.5468807	0.5787523	0.7709150	-0.7093917	0.4780814	0.0920419
ChrRegiontelomeric	1.1484696	3.1533632	0.7468320	1.5377883	0.1241004	0.2172661
Length.Mb.Scaled	-0.0378516	0.9628558	0.3111212	-0.1216620	0.9031667	0.1329224
WAvgRate.perMb.Scaled	-0.5982169	0.5497911	0.4530147	-1.3205244	0.1866600	0.0641060

4 Possible improvements / To do list

It would be nice to calculate power and goodness of fit for these models, especially the NAHR one, which is the one that needs more caution on interpretation.