Detection of large scale inversion location patterns with ordinal logistic regression

Ruth Gómez Graciani

0.1 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 1 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.

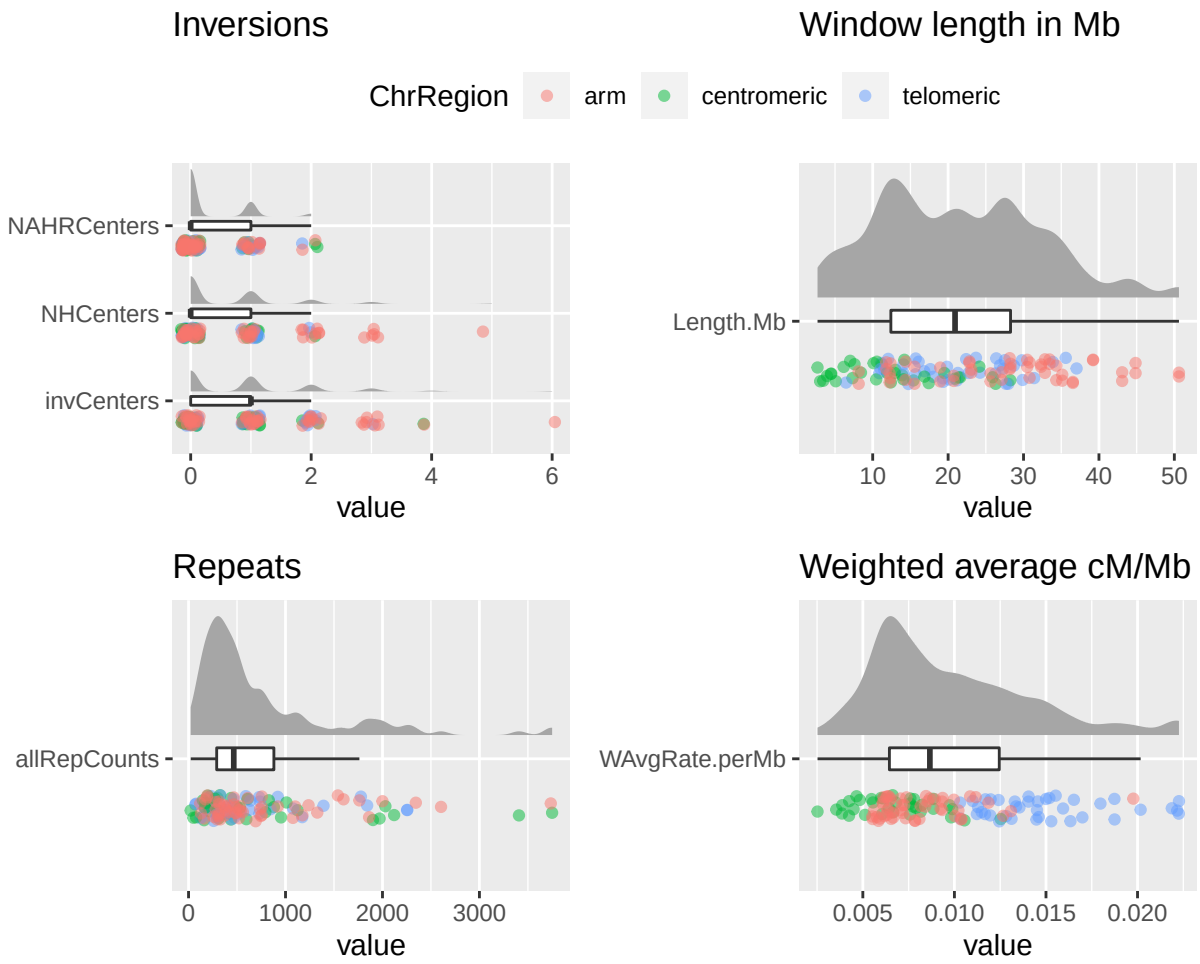

Figure 1: Raincloud plots for each variable.

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 2.

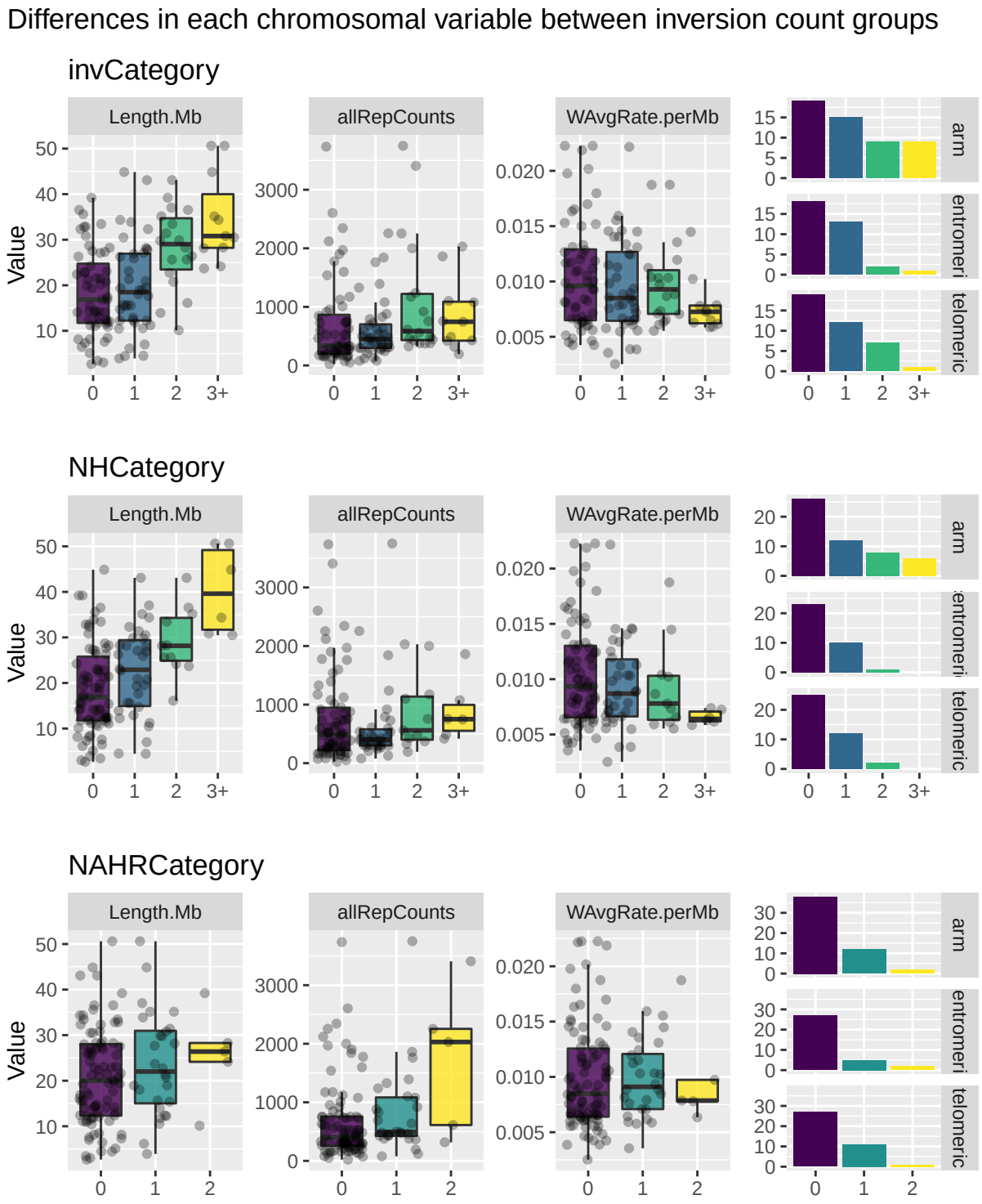

Figure 2: Boxplots for each dependent variable gorup and each independent variable quickly show candidates of having a strong effect. We can also see that there is missing data for some chromosome region types, because windows with 3+ inversions are scarce.

Finally, I will test assumption number 3, no multi-collinearity between independent variables. Figure 3 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.

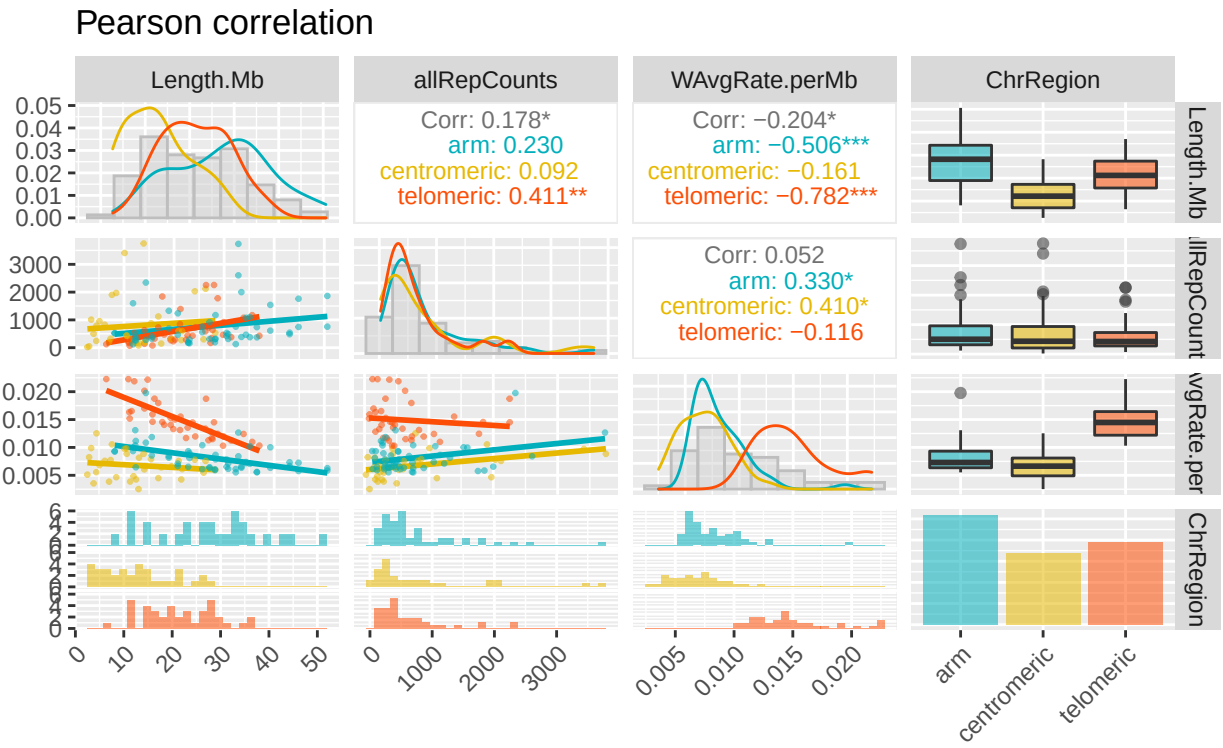

Figure 3: Correlation matrix of independent variables.

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.

0.2 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