Detection of large scale inversion location patterns with ordinal logistic regression

Ruth Gómez Graciani

Contents

Contents

1 Windows generation

In this report, I analyze the realtionship between the amount of chromosomes in heterozygosis and amount of aneuploidies, which is key to distinguish whether the observed reduction in crossover rate in heterozygotes is due to a physical impediment of recombination, or due to a generation of aberrations and later discard from the crossover dataset.

2 Introduction

There are two possible mechanisms that can lead to recombination reduction between opposite orientations of an inversion:

- A physical impediment to pair or recombine.
- Purifying selection against the unbalanced results of a crossover between orientations.

In Bell et al. (2020), they sequence 900 to 2000 gametes from 20 donors, and then the sequenced chromosomes have 3 possible fates:

- Halthy chromosome: goes to the crossover dataset and it is used to elaborate single-individual crossover maps.
- Chromosome with small and medium aberrations: it is discarded.
- Whole chromosome and chromosome arm gains and losses: they are counted in the aberration dataset.

We observe a reduction in crossover rate in heterozygotes compared to homozygotes, but due to the methodology in Bell et al. 2020, this is not enough to confirm whether the inhibition mechanism is physical inhibition or purifying selection. However, we can complement that observation with the measurement of alterations in the aberration count caused by inversions in heterozygosis. If more heterozygous inversions leads to more aberrations in that chromosome, that would be consistent with a purifying selection scenario.

3 Variables to test

As per the independent variable, I initially measured, on each chromosome and individual, how many centiMorgans were affected by inversions in heterozygosis as a proxy of the probability of having an aberration-generating crossover. In addition, I selected some inversion subsets that may have a larger impact in the aberration count:

- Inversions near the centromere will form a very large acentric spanning almost all the chromosome arm, so we expect them to influence the number of chromosome arm losses. I counted as centromeric regions the 20% of each arm next to the centromere, since any inversion in that region would cause an acentric of at least ~80% of the arm.
- Inversions near the telomere will form a very small acentric fragment, but if the dicentric fragment breaks near the centromere, the results of the crossover will be a chromosome arm gain and a chromosome arm loss. In addition, these regions concentrate most of the crossovers in the cromosome. I counted as telomeric regions the same regions we selected based on recombination density in the population-based recombination map analysis.
- Big inversions represent a highest chance of crossover just by probability, and we observed a consistent reduction in crossovers in heterozygotes compared to homozygotes. All those discarded crossovers could be contributing to the aberration count. I considered as big inversions those that span more than 1 window in the single-individual crossover maps (>200 kb).

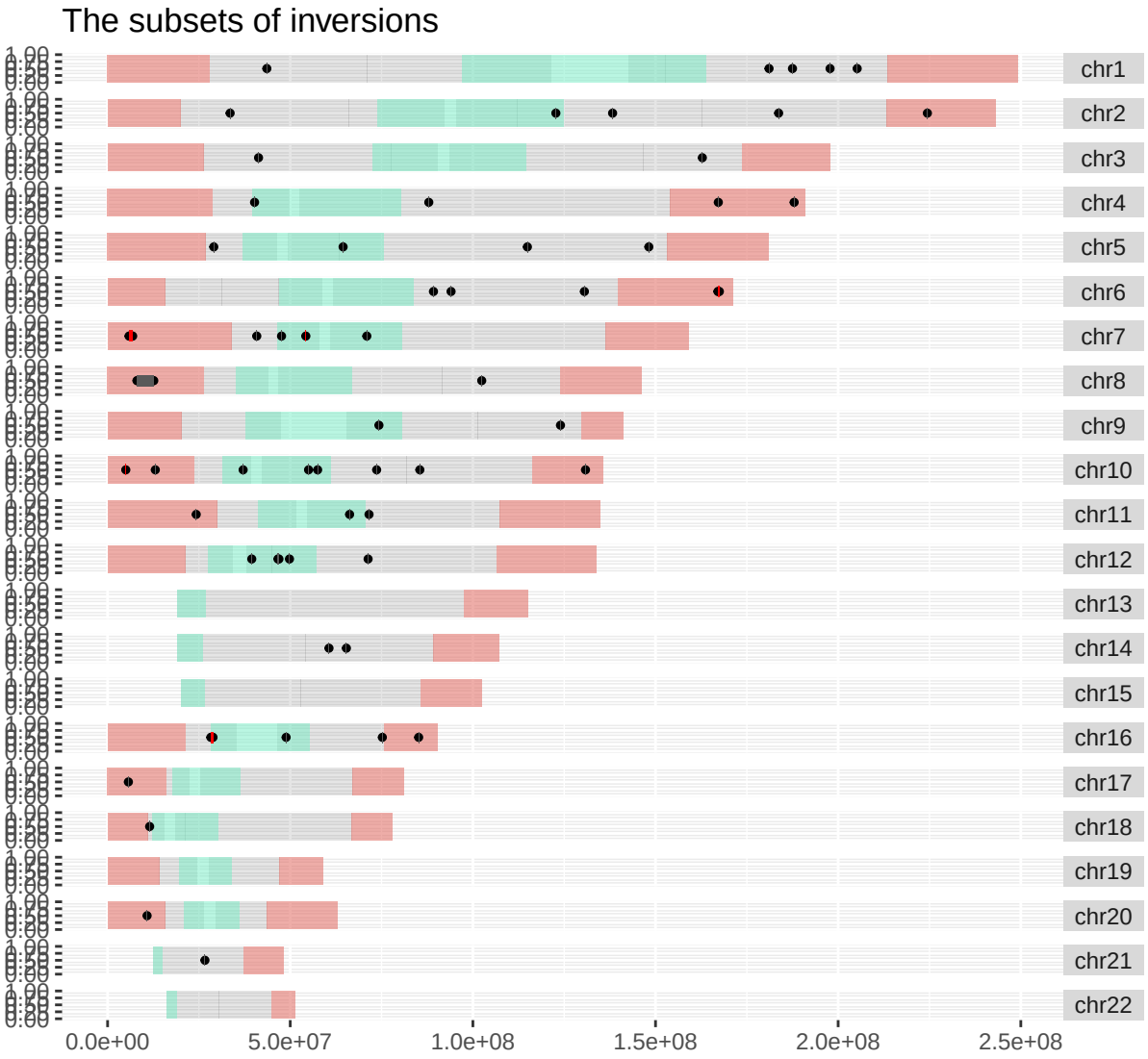

Figure 1: The selection of inversion subsets. The map windows that contain the inversion are marked in as dark gray lines (or red for big inversions), and also with points for easier visualization. Inversions within the green region are considered to be near the centromere. Inversions within the blue region are considered to be near the telomere.

As a dependent variable, I used the number of arm gains and arm losses, which are the types of aberrations expected to be influenced by the independent variable.

In our fist attempts in this analysis I made some correlations, however, as it will be shown later, a big portion of cases had 0 heterozygous inversions and/or 0 aberrations detected, so I opted for converting them into binomial variables and comparing absence and presence of each.

4 Data preparation

4.1 Model assumptions

My data points are one measurement for each chromosome and individual, and for them I have information about crossover rate, number of cells, heterozygous inversions, number of different types of aneluploidies. Telocentric chromosomes should be discarded from the arm aberration measurements, since they were counted as whole chromosome aberrations.

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 3.

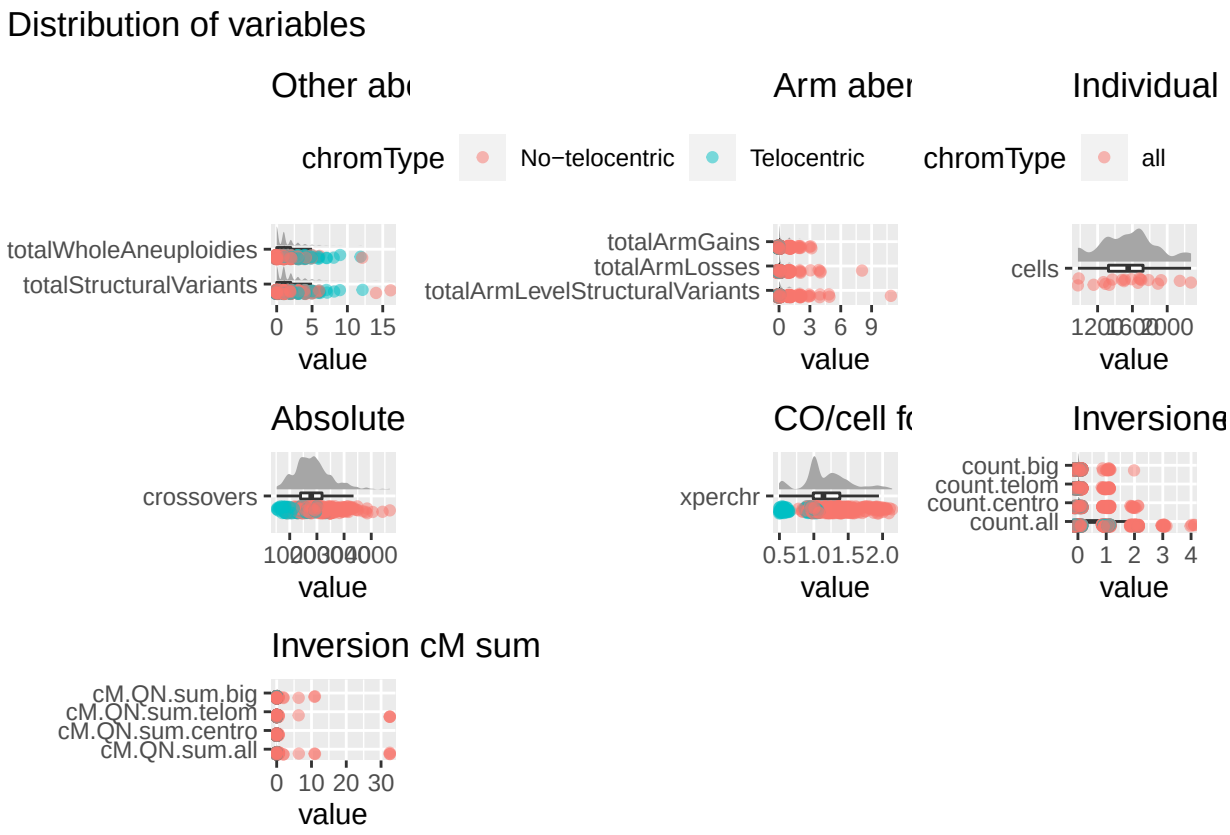

Figure 2: Raincloud plots for each variable.

From now on, no telocentric chroms

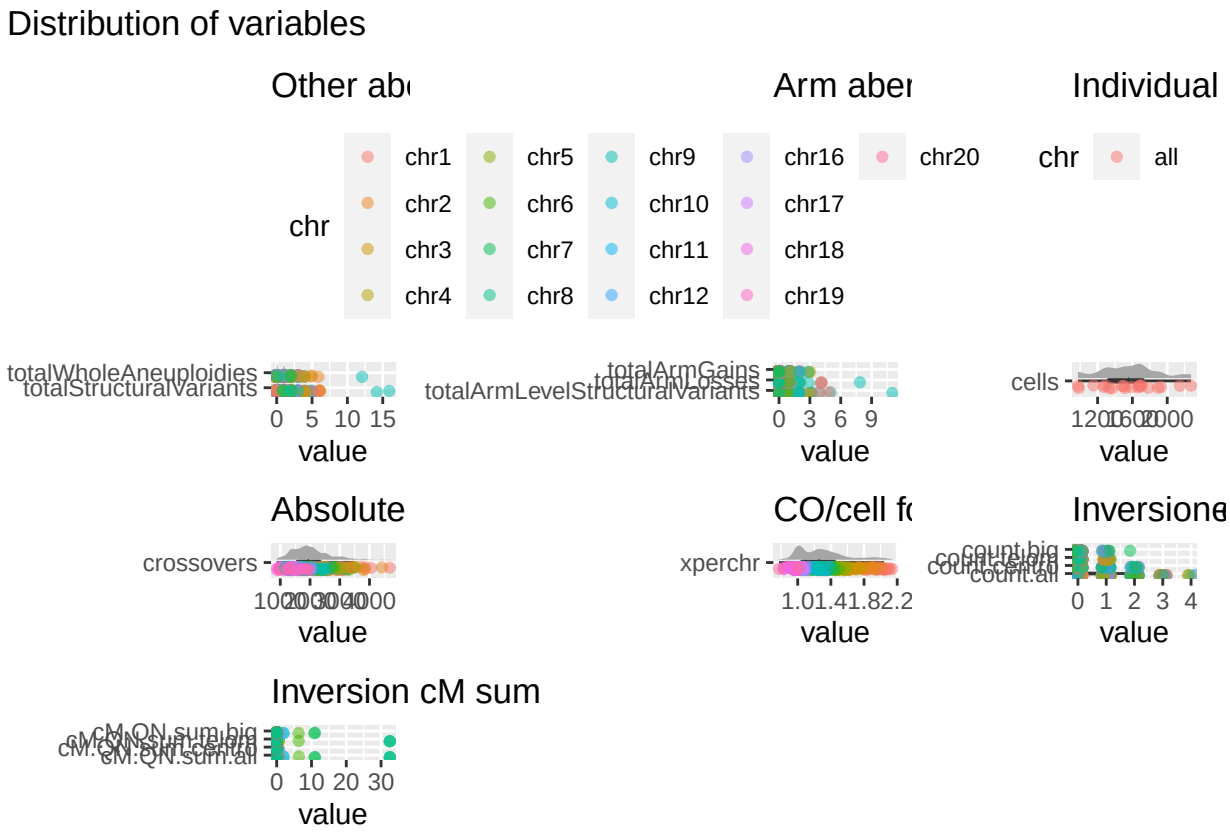

Figure 3: Raincloud plots for each variable.

Table 1: Original category counts

CountGroups	totalStructuralVariants	totalWholeAneuploidies	totalArmLevelStructuralVariants	totalArmLosses	
0	132	171	258	300	
1	112	108	58	31	
11	NA	NA	1	NA	
12	NA	1	NA	NA	
14	1	NA	NA	NA	
16	1	NA	NA	NA	
2	55	39	15	4	
3	20	14	4	1	
4	10	4	2	3	
5	5	2	2	NA	
6	4	1	NA	NA	
8	NA	NA	NA	1	

Table 2: New category counts

CountGroups	totalStructuralVariantsCategory	totalWholeAneuploidiesCategory	totalArmLevelStructuralVariantsCa
0	132	171	
1	112	108	
2	55	39	
3+	41	22	

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

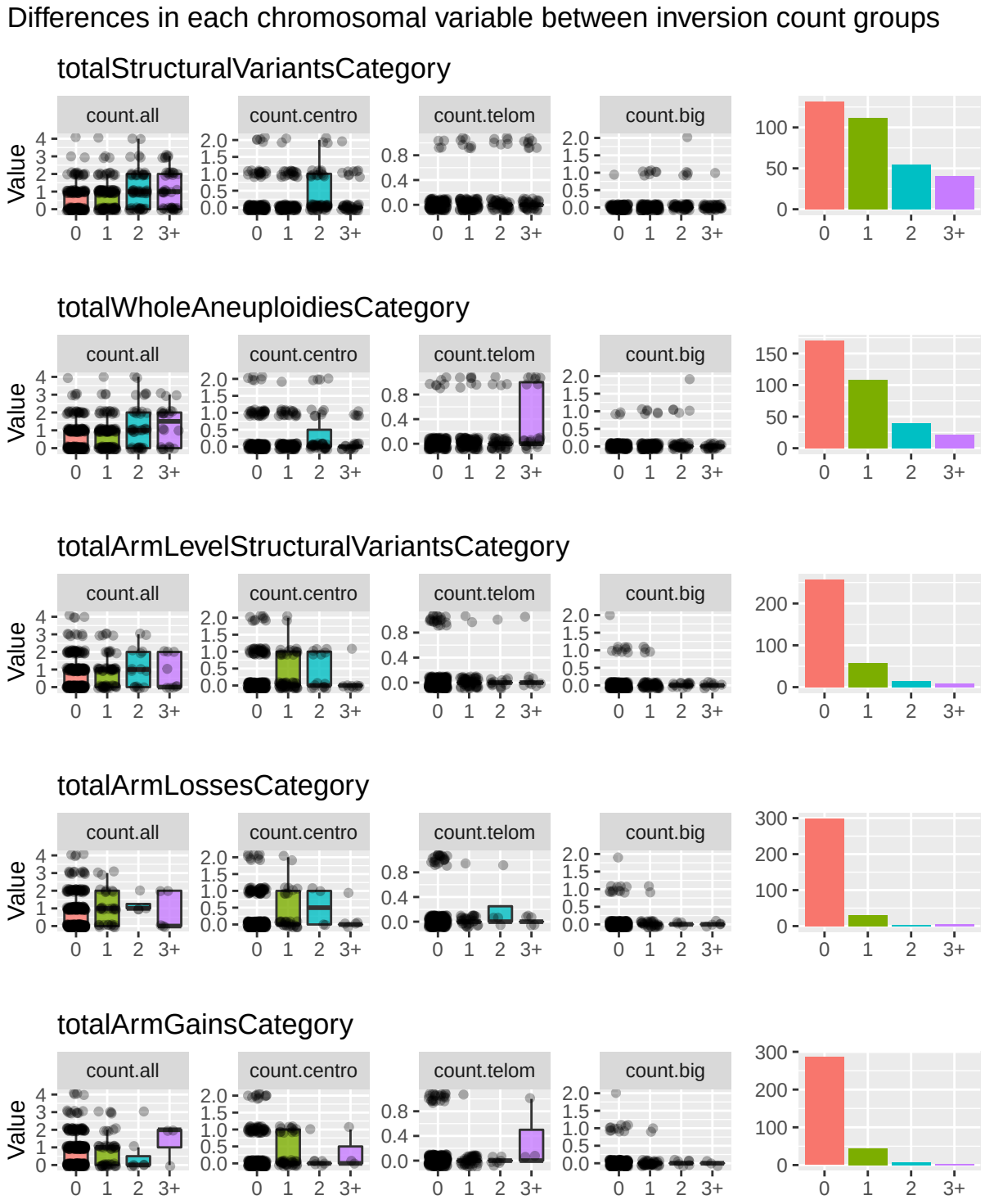

Figure 4: Boxplots for each dependent variable group and each independent variable quickly show candidates of having a strong effect. We can also see that there is missing data for some chromosome region types, because windows with 3+ inversions are scarce.

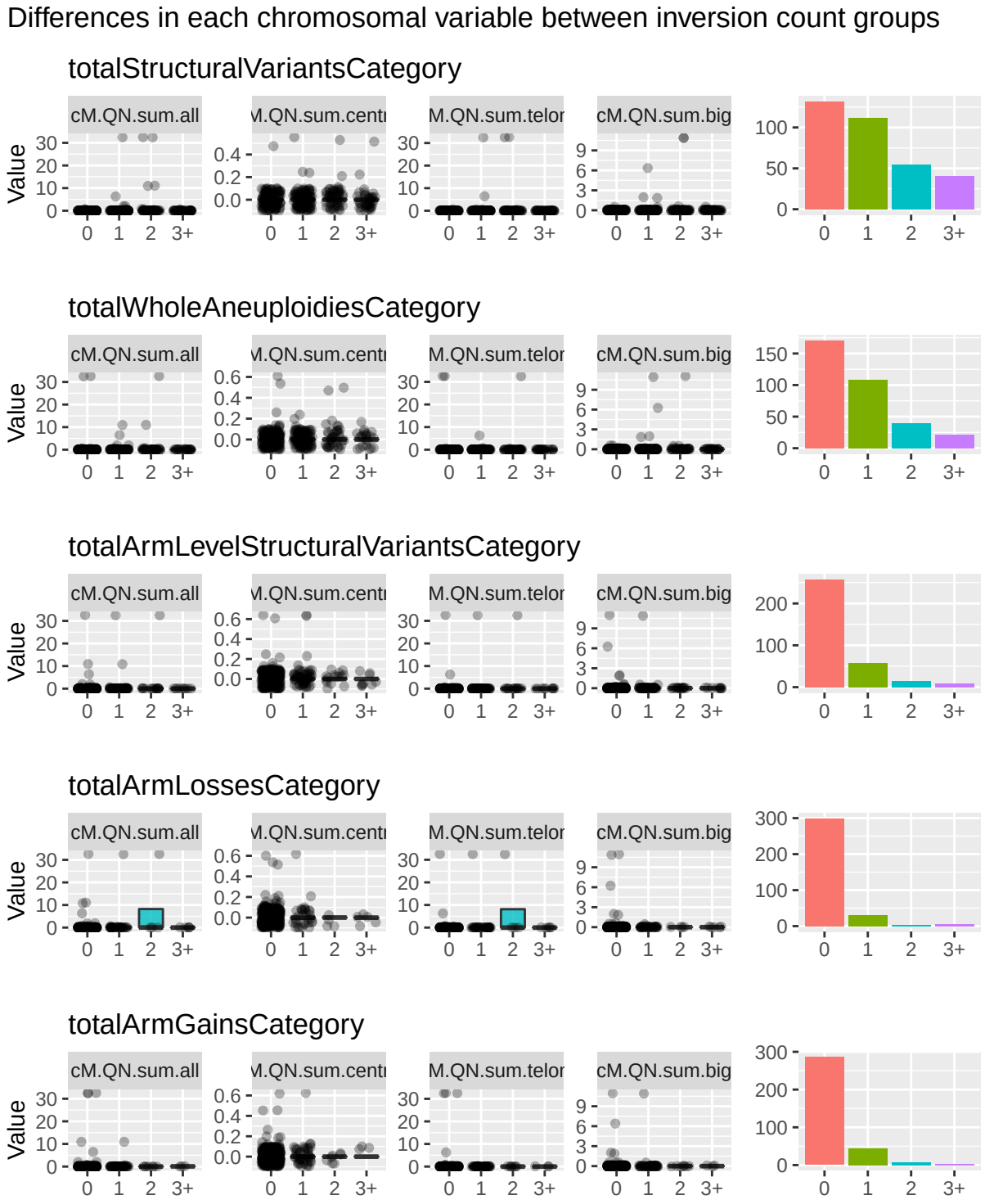

Figure 5: Boxplots for each dependent variable group and each independent variable quickly show candidates of having a strong effect. We can also see that there is missing data for some chromosome region types, because windows with 3+ inversions are scarce.

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 6). 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.

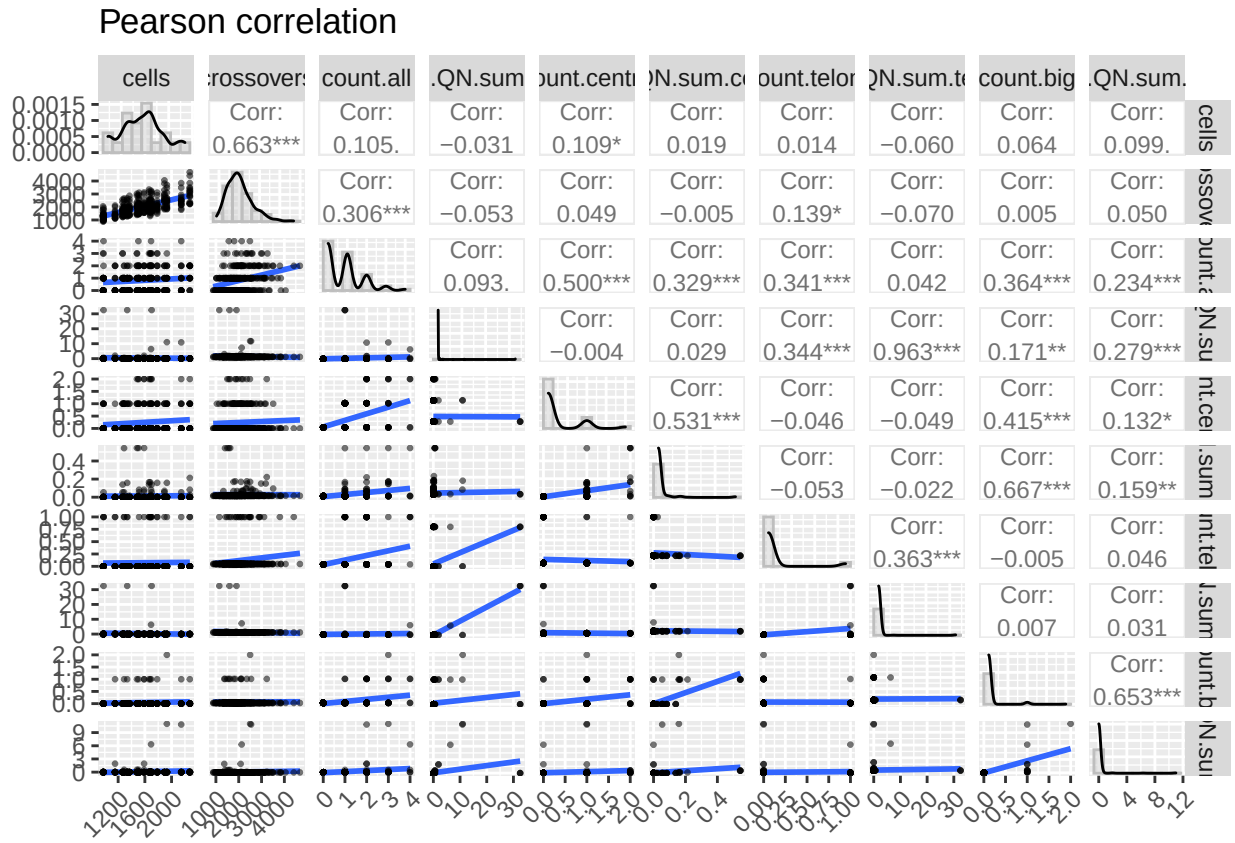

Figure 6: Correlation matrix of independent variables.

Table 3: Variance Inflation Factor

	x
cells	1.838675
crossovers	2.036542
count.all	1.154958
cM.QN.sum.all	1.017022

Table 4: Variance Inflation Factor

	x
cells	1.823457
crossovers	1.887041
count.telom	1.207075
cM.QN.sum.telom	1.173142

Table 5: Variance Inflation Factor

	x
cells	1.817499
crossovers	1.801033
count.centro	1.418274
cM.QN.sum.centro	1.401721

Table 6: Variance Inflation Factor

	x
cells	1.815864
crossovers	1.804669
count.big	1.746590
cM.QN.sum.big	1.753312

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

4.2 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.

5 Model fitting

5.1 All inversions

5.1.1 All Structural Variants

Table 7: Model summary for totalStructuralVariantsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.2383314	1.269130	0.1366633	1.743931	0.0811712	0.2205023
cM.QN.sum.all.scaled	0.1457768	1.156938	0.0832356	1.751376	0.0798812	0.1912057
count.all.scaled	0.2578427	1.294135	0.1107688	2.327755	0.0199251	0.2859947
crossovers.scaled	0.3115401	1.365527	0.1424781	2.186582	0.0287731	0.5937673

5.1.2 Whole Aneuploidies

Table 8: Model summary for totalWholeAneuploidiesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.3213536	1.378993	0.1398728	2.2974711	0.0215919	0.2072841
cM.QN.sum.all.scaled	0.0154742	1.015594	0.1140179	0.1357171	0.8920449	0.4627604
count.all.scaled	0.2857592	1.330772	0.1141733	2.5028547	0.0123196	0.3630562
crossovers.scaled	0.1533898	1.165779	0.1438378	1.0664086	0.2862390	0.6476569

5.1.3 All Arm Aberrations

Table 9: Model summary for totalArmLevelStructuralVariantsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.0405467	1.041380	0.1705323	0.2377657	0.8120629	0.4815033
cM.QN.sum.all.scaled	0.1913389	1.210870	0.0966245	1.9802325	0.0476774	0.9798580
count.all.scaled	0.0205825	1.020796	0.1341842	0.1533897	0.8780910	0.6781640
crossovers.scaled	0.2978077	1.346903	0.1709572	1.7420011	0.0815082	0.2342369

5.1.4 Arm Losses

Table 10: Model summary for totalArmLossesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.0251054	1.025423	0.2272405	0.1104795	0.9120291	0.7520208
cM.QN.sum.all.scaled	0.2564643	1.292353	0.1015532	2.5254180	0.0115561	0.6138944
count.all.scaled	0.1425070	1.153161	0.1695904	0.8403009	0.4007397	0.6705981
crossovers.scaled	0.2236651	1.250652	0.2255533	0.9916287	0.3213787	0.2178509

5.1.5 Arm Gains

Table 11: Model summary for totalArmGainsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.0031834	1.0031885	0.1982413	0.0160584	0.9871878	0.4844172
cM.QN.sum.all.scaled	-0.0637007	0.9382858	0.2092780	-0.3043834	0.7608358	0.9725800
count.all.scaled	-0.0850952	0.9184249	0.1615574	-0.5267176	0.5983897	0.2953711
crossovers.scaled	0.3542837	1.4251595	0.1888411	1.8760946	0.0606423	0.3386680

5.2 Centromeric inversions

5.2.1 Whole Aneuploidies

Table 12: Model summary for totalWholeAneuploidiesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.2774562	1.3197684	0.1383975	2.0047782	0.0449868	0.1329775
cM.QN.sum.centro.scaled	0.1951711	1.2155190	0.1250554	1.5606773	0.1185999	0.1958467
count.centro.scaled	-0.1040048	0.9012209	0.1260486	-0.8251171	0.4093051	0.6844932
crossovers.scaled	0.2722403	1.3129024	0.1352753	2.0124902	0.0441683	0.3852386

5.2.2 Arm Losses

```
## [1] "Using totalArmLossesCategory~cells.scaled + crossovers.scaled + count.all.scaled + cM.QN.sum.al
```

Table 13: Model summary for totalArmLossesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	-0.0500937	0.9511403	0.2300915	-0.2177121	0.8276535	0.8563308
cM.QN.sum.centro.scaled	-0.0188622	0.9813146	0.1751375	-0.1076994	0.9142341	0.6766923
count.centro.scaled	0.2173627	1.2427948	0.1775189	1.2244485	0.2207831	0.2995013
crossovers.scaled	0.2745939	1.3159961	0.2165964	1.2677676	0.2048810	0.2636345

5.2.3 Arm Gains

```
## [1] "Using totalArmGainsCategory~cells.scaled + crossovers.scaled + count.all.scaled + cM.QN.sum.all
```

Table 14: Model summary for totalArmGainsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.0072366	1.007263	0.1989815	0.0363682	0.9709888	0.4934138
cM.QN.sum.centro.scaled	0.0871080	1.091014	0.1545906	0.5634752	0.5731113	0.8863411
count.centro.scaled	0.0199293	1.020129	0.1732263	0.1150476	0.9084074	0.6914396
crossovers.scaled	0.3340183	1.396569	0.1812643	1.8427144	0.0653707	0.4816830

5.3 Telomeric inversions

5.3.1 Whole Aneuploidies

Table 15: Model summary for totalWholeAneuploidiesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.3510876	1.4206118	0.1410973	2.488266	0.0128367	0.2466270
cM.QN.sum.telom.scaled	-0.1501779	0.8605548	0.1302878	-1.152663	0.2490488	0.1782803
count.telom.scaled	0.4209398	1.5233926	0.1175577	3.580709	0.0003427	0.0463702
crossovers.scaled	0.1673755	1.1821982	0.1382882	1.210338	0.2261491	0.7363790

5.3.2 Arm Losses

```
## [1] "Using totalArmLossesCategory~cells.scaled + crossovers.scaled + count.all.scaled + cM.QN.sum.al
## Error in optim(s0, fmin, gmin, method = "BFGS", ...): initial value in 'vmmin' is not finite
```

Table 16: Model summary for totalArmLossesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	-0.0245719	0.9757276	0.2289387	-0.1073295	0.9145276	0.7221085
cM.QN.sum.telom.scaled	1.5427810	4.6775807	30.9470805	0.0498522	0.9602401	1.0000000
count.telom.scaled	-3.5123676	0.0298262	85.8807772	-0.0408982	0.9673771	1.0000000
crossovers.scaled	0.3761764	1.4567041	0.2158359	1.7428814	0.0813543	0.3589159

5.3.3 Arm Gains

Table 17: Model summary for totalArmGainsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	-0.0129858	0.9870982	0.1969028	-6.595010e-02	0.9474176	0.5259157
cM.QN.sum.telom.scaled	-1663.0980718	0.0000000	0.0158733	-1.047735e+05	0.0000000	1.0000000
count.telom.scaled	0.8232635	2.2779216	0.2079892	3.958202e+00	0.0000755	0.9652963
crossovers.scaled	0.3618406	1.4359701	0.1833654	1.973331e+00	0.0484578	0.3014387

5.4 Big inversions

5.4.1 Whole Aneuploidies

Table 18: Model summary for totalWholeAneuploidiesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.2520495	1.2866597	0.1385176	1.8196202	0.0688169	0.1897250
cM.QN.sum.big.scaled	-0.0379142	0.9627956	0.1147150	-0.3305074	0.7410166	0.6589363
count.big.scaled	0.2162421	1.2414029	0.1236425	1.7489305	0.0803030	0.5489111
crossovers.scaled	0.2834260	1.3276707	0.1358791	2.0858695	0.0369904	0.4790617

5.4.2 Arm Losses

Table 19: Model summary for totalArmLossesCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	-0.0094750	0.9905698	0.2264481	-0.0418417	0.9666249	0.7863866
cM.QN.sum.big.scaled	-0.9465842	0.3880643	1.7478853	-0.5415597	0.5881219	0.9999982
count.big.scaled	0.2227683	1.2495311	0.2388853	0.9325328	0.3510612	0.9999828
crossovers.scaled	0.2544685	1.2897760	0.2108824	1.2066844	0.2275537	0.3906267

5.4.3 Arm Gains

Table 20: Model summary for totalArmGainsCategory

Variable	log(OddsRatio)	OddsRatio	Std.Error	t.value	p.value	Brant p.value
cells.scaled	0.0024990	1.0025021	0.1974114	0.0126589	0.9898999	0.4449830
cM.QN.sum.big.scaled	0.2206077	1.2468342	0.1838162	1.2001537	0.2300796	1.0000000
count.big.scaled	-0.2590820	0.7717597	0.2736966	-0.9466029	0.3438412	0.9999915
crossovers.scaled	0.3249697	1.3839888	0.1804621	1.8007645	0.0717400	0.4055255