Ordinal logistic model on large, classified windows data from Spence

Ruth Gómez Graciani

Prepare the data

First, we obtain the density distribution, and local minima and maxima for the recombination map.

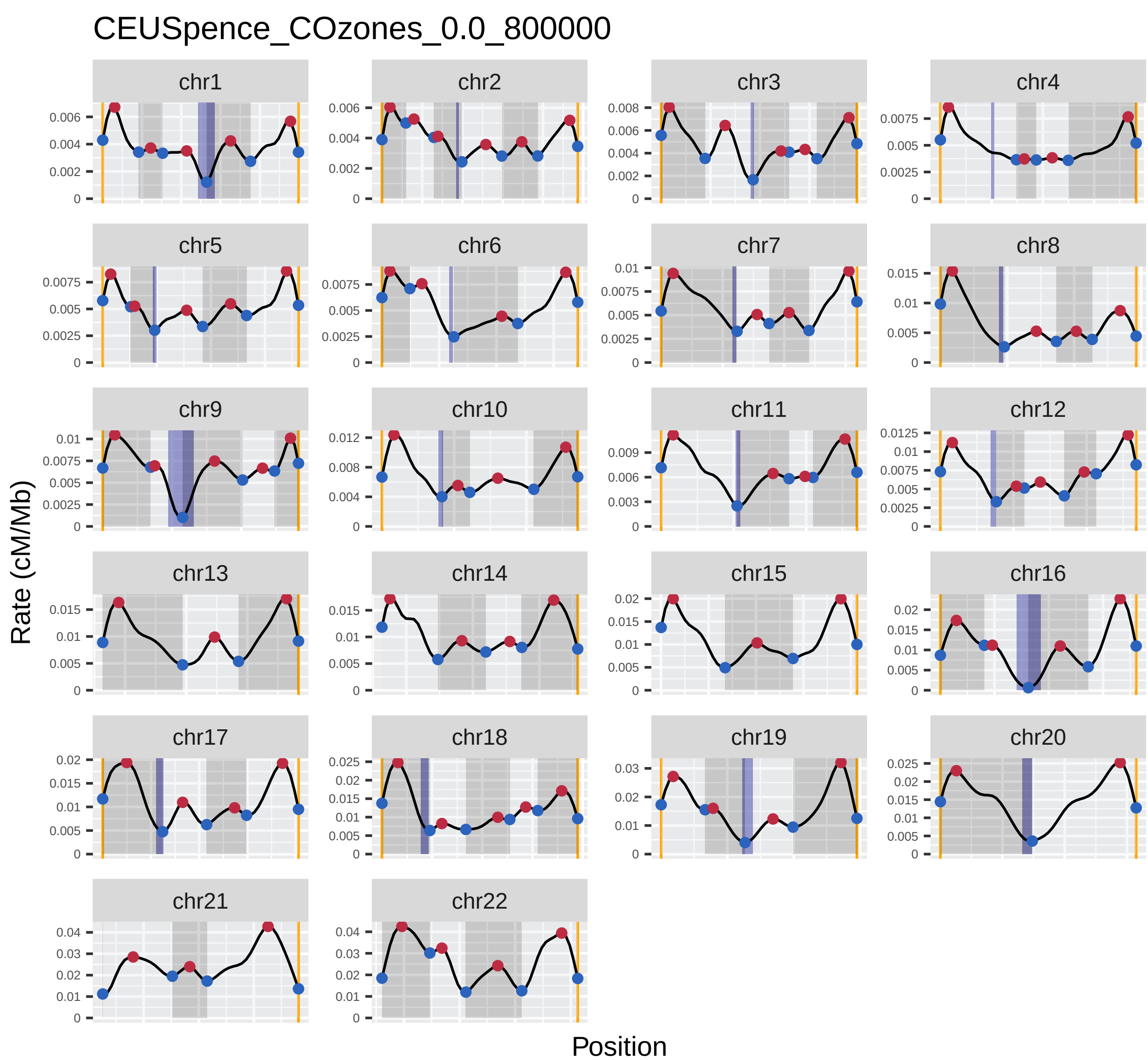


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

Next, we define telomeric regions as the space between the chromosome start to the next local minimum, or between the chromosome end to the previous local minimum. We also define the limits of the centromere by calculating the midpoint between the flanking maxima and the local minimum near the centromere (or the centromere itself if there is no local minimum). Actual centromeres will be excluded from centromeric regions to avoid biases, specially in the Spence data. These categories will be represented as the “Color” variable in the statistical analysis.

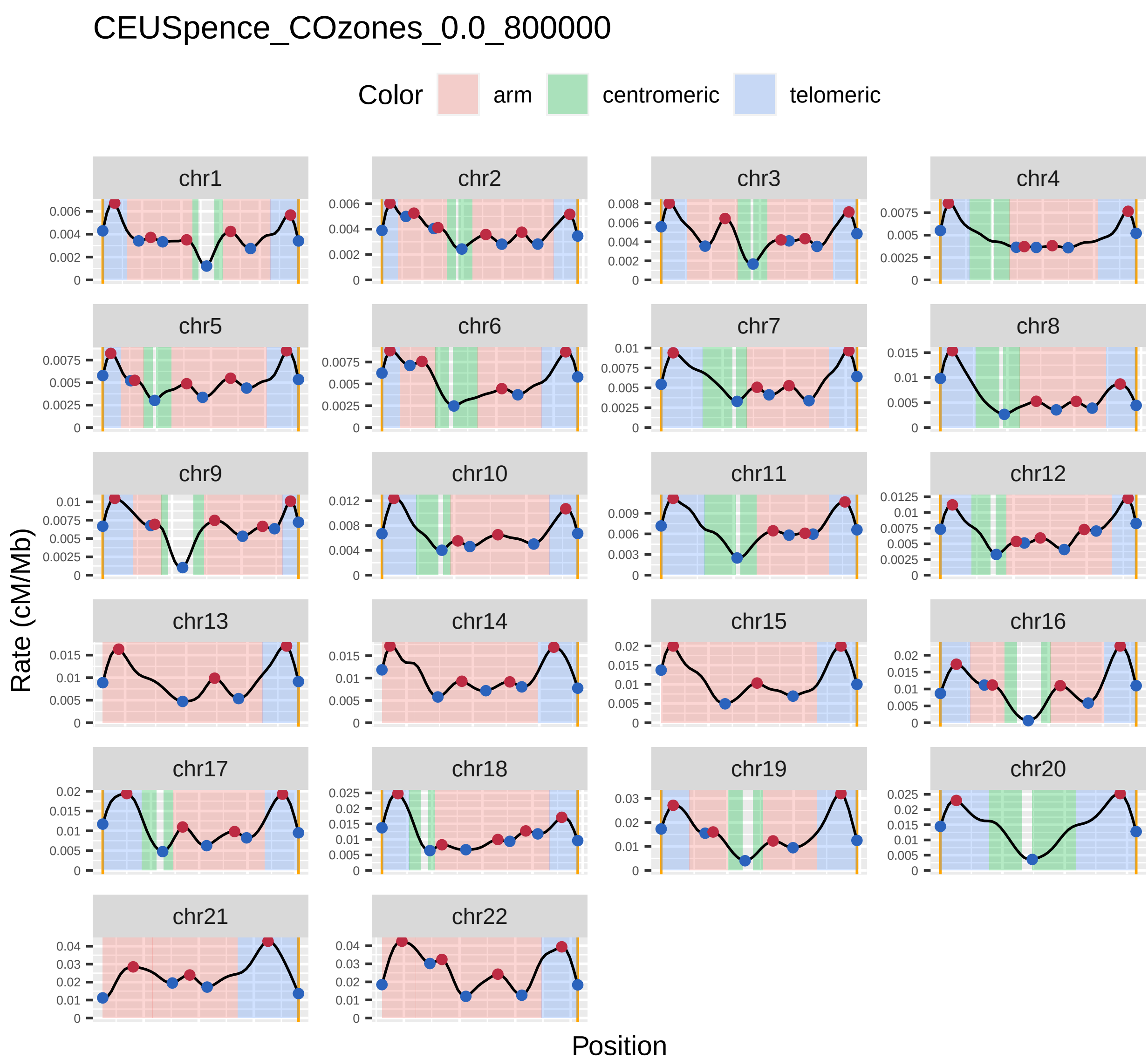


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Descriptive statistics

Raw data:

```
##    Chromosome     Start       End       Color invCenters NHCenters NAHRCenters
## 1       chr10     60683  23750219   telomeric          3         2           1
## 2       chr10  23750219  39146059 centromeric          1         0           1
## 3       chr10  47478464 116172416         arm          4         4           0
## 4       chr10 116172416 135524372   telomeric          1         1           0
## 5       chr10  42369508  47478464 centromeric          0         0           0
## 6       chr11     87267  29960849   telomeric          2         1           1
##    Length.Mb allRepCounts WAvgRate.perMb
## 1 23.689536          428    0.014480534
## 2 15.395840          880    0.007907887
## 3 68.693953         1542    0.007285414
## 4 19.351956          266    0.014043434
## 5  5.108956          954    0.006810256
## 6 29.873582          920    0.011957306
```

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 on different subsets of the data. 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 below. The inversion counts have only a number of possible options, so they can be considered an ordinal variable. The independent variables are continuous and categorical, so assumptions 1 and 2 are satisfied

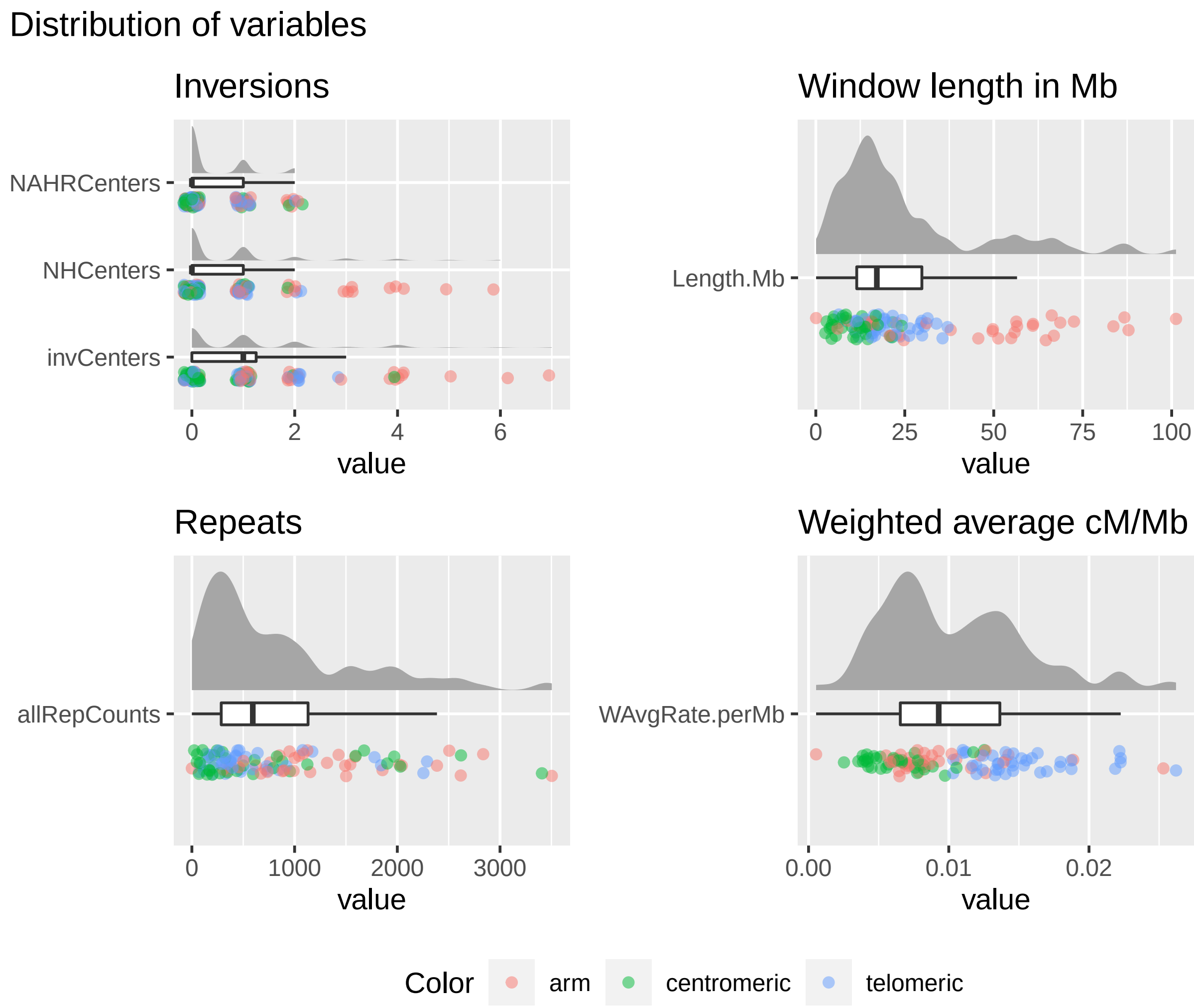


Figure 3: Distribution of variables.

We see that some categories have low number of cases, so I will make a "3 or more" category when relevant.

```
## [1] "Original counts"

##   CountGroups invCenters NHCenters NAHRCenters
## 1           0         49        65          78
## 2           1         32        27          22
## 3           2         15         7           8
## 4           3          2         4          NA
## 5           4          7         3          NA
## 6           5          1         1          NA
## 7           6          1         1          NA
## 8           7          1        NA          NA

## [1] "New counts"

##   CountGroups invCategory NHCategory NAHRCategory
## 1           0          49         65           78
## 2           1          32         27           22
## 3           2          15          7            8
## 4          3+          12          9           NA
```

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

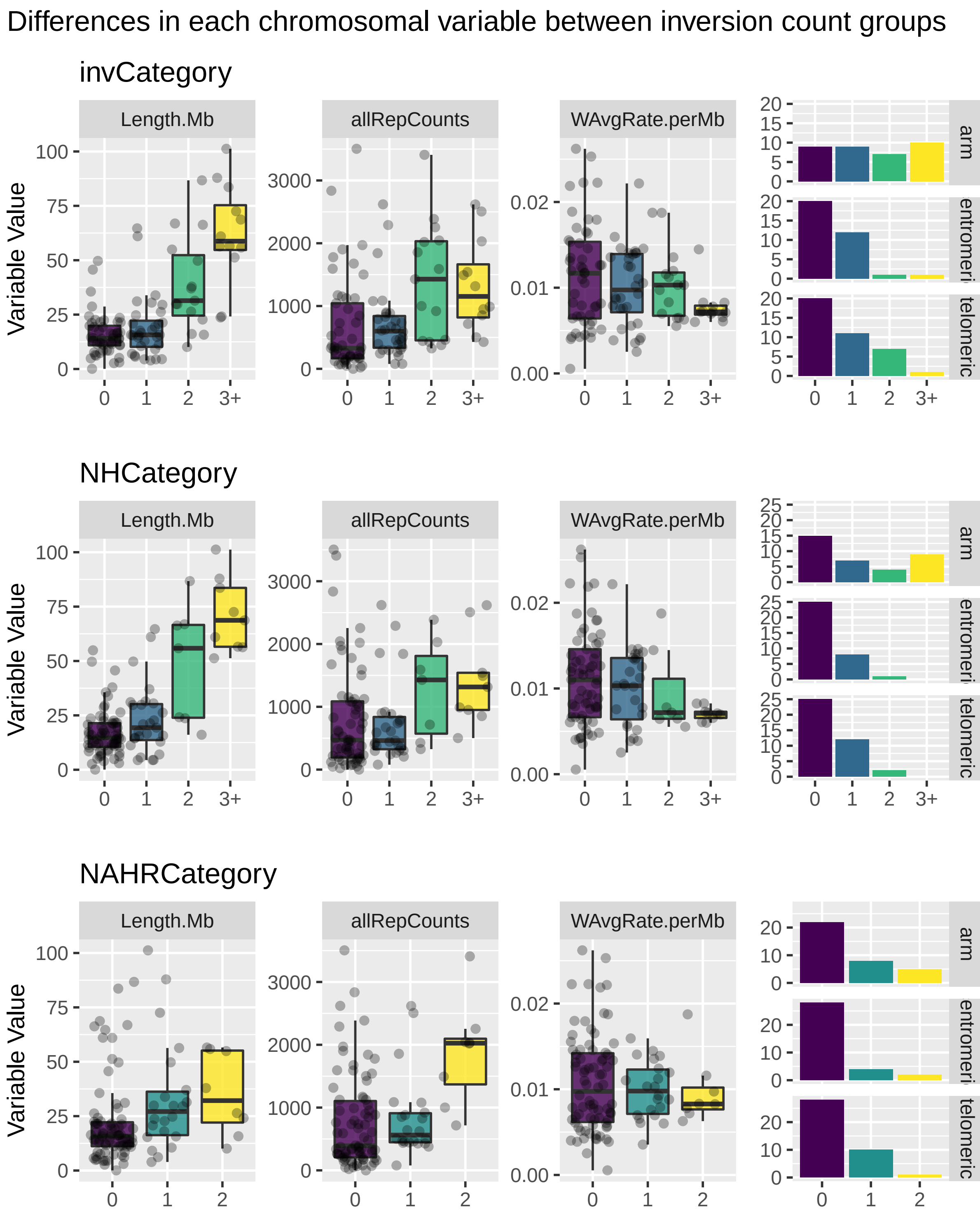


Figure 4: Potential effect of independent variables on the different types of invesions.

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

Figure 5: Correlations between variables.

We see that our three variables are significantly correlated, but this does not confirm multi-collinearity. I perform a variance inflation factor test on the corresponging linear model to further check the multi-collinearity.

```
##                     GVIF Df GVIF^(1/(2*Df))
## Length.Mb       2.554186  1        1.598182
## allRepCounts    1.329840  1        1.153187
## Color           5.084803  2        1.501649
## WAvgRate.perMb  3.040313  1        1.743649
```

The general rule of thumbs for VIF test is that if the VIF value is greater than 10, then there is multi-collinearity, so we can say that the third assumption (no multi-collinearity) is satisfied.

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

Variable scalation (optional)

Standardized coefficients are useful in our case to compare effects of predictors reported in different units. The most straightforward way is using the Agresti method of standardization, applied with the `scale()` function.

```
##    Length.Mb         Length.Mb.Scaled   allRepCounts    allRepCounts.Scaled
##  Min.   :  0.07717   Min.   :-1.1754   Min.   :   0.0   Min.   :-1.0881
##  1st Qu.: 11.49431   1st Qu.:-0.6367   1st Qu.: 285.5   1st Qu.:-0.7244
##  Median : 17.14259   Median :-0.3701   Median : 593.0   Median :-0.3326
##  Mean   : 24.98653   Mean   : 0.0000   Mean   : 854.1   Mean   : 0.0000
##  3rd Qu.: 29.80087   3rd Qu.: 0.2272   3rd Qu.:1131.0   3rd Qu.: 0.3528
##  Max.   :101.22393   Max.   : 3.5975   Max.   :3504.0   Max.   : 3.3761
##  WAvgRate.perMb      WAvgRate.perMb.Scaled
##  Min.   :0.0005371   Min.   :-1.9171
##  1st Qu.:0.0065417   1st Qu.:-0.7589
##  Median :0.0092825   Median :-0.2302
##  Mean   :0.0104759   Mean   : 0.0000
##  3rd Qu.:0.0136431   3rd Qu.: 0.6109
##  Max.   :0.0262073   Max.   : 3.0344
```

Once the model is fitted, we can use the sd to transform scaled coefficients to natural coefficients and viceversa.

Total inversions (invCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error    t value
## Length.Mb         6.453e-02   0.014794     4.3622
## allRepCounts      3.558e-04   0.000305     1.1666
## Colorcentromeric -1.729e-01   0.612130    -0.2825
## Colortelomeric    1.824e-01   0.530789     0.3436
## WAvgRate.perMb   -5.503e+01   0.010046 -5477.1084
##
## Intercepts:
##      Value      Std. Error t value
## 0|1      0.8257     0.6573     1.2562
## 1|2      2.6771     0.7307     3.6639
## 2|3+     4.3539     0.8677     5.0176
##
## Residual Deviance: 214.4465
## AIC: 230.4465
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                          Value   Std. Error       t value    p value
## Length.Mb         6.453460e-02 0.0147941536     4.3621692 0.00001288
## allRepCounts      3.557639e-04 0.0003049553     1.1666099 0.24336795
## Colorcentromeric -1.729144e-01 0.6121297167    -0.2824799 0.77757554
## Colortelomeric    1.823657e-01 0.5307890421     0.3435747 0.73116614
## WAvgRate.perMb   -5.502506e+01 0.0100463714 -5477.1083908 0.00000000
## 0|1               8.256780e-01 0.6572867597     1.2561914 0.20904658
## 1|2               2.677078e+00 0.7306609046     3.6639128 0.00024839
## 2|3+              4.353896e+00 0.8677297388     5.0175711 0.00000052
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                          2.5 %       97.5 %
## Length.Mb         0.0331935576 0.0997318829
## allRepCounts     -0.0002242382 0.0009214449
## Colorcentromeric -1.6078287381 1.3207436277
## Colortelomeric   -0.9887336089 1.3806044999
## WAvgRate.perMb              NA           NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                          2.5 %        97.5 %
## Length.Mb         3.553859e-02  9.353061e-02
## allRepCounts     -2.419375e-04  9.534652e-04
## Colorcentromeric -1.372667e+00  1.026838e+00
## Colortelomeric   -8.579617e-01  1.222693e+00
```

```
## WAvgRate.perMb   -5.504476e+01 -5.500537e+01
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                    Odds Ratio      2.5%    97.5%
## Length.Mb        1.066662e+00 1.0337506 1.104875
## allRepCounts     1.000356e+00 0.9997758 1.000922
## Colorcentromeric 8.412097e-01 0.2003221 3.746206
## Colortelomeric   1.200053e+00 0.3720476 3.977305
## WAvgRate.perMb   1.267412e-24        NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0666625 times more likely to increase in inversion amount category."

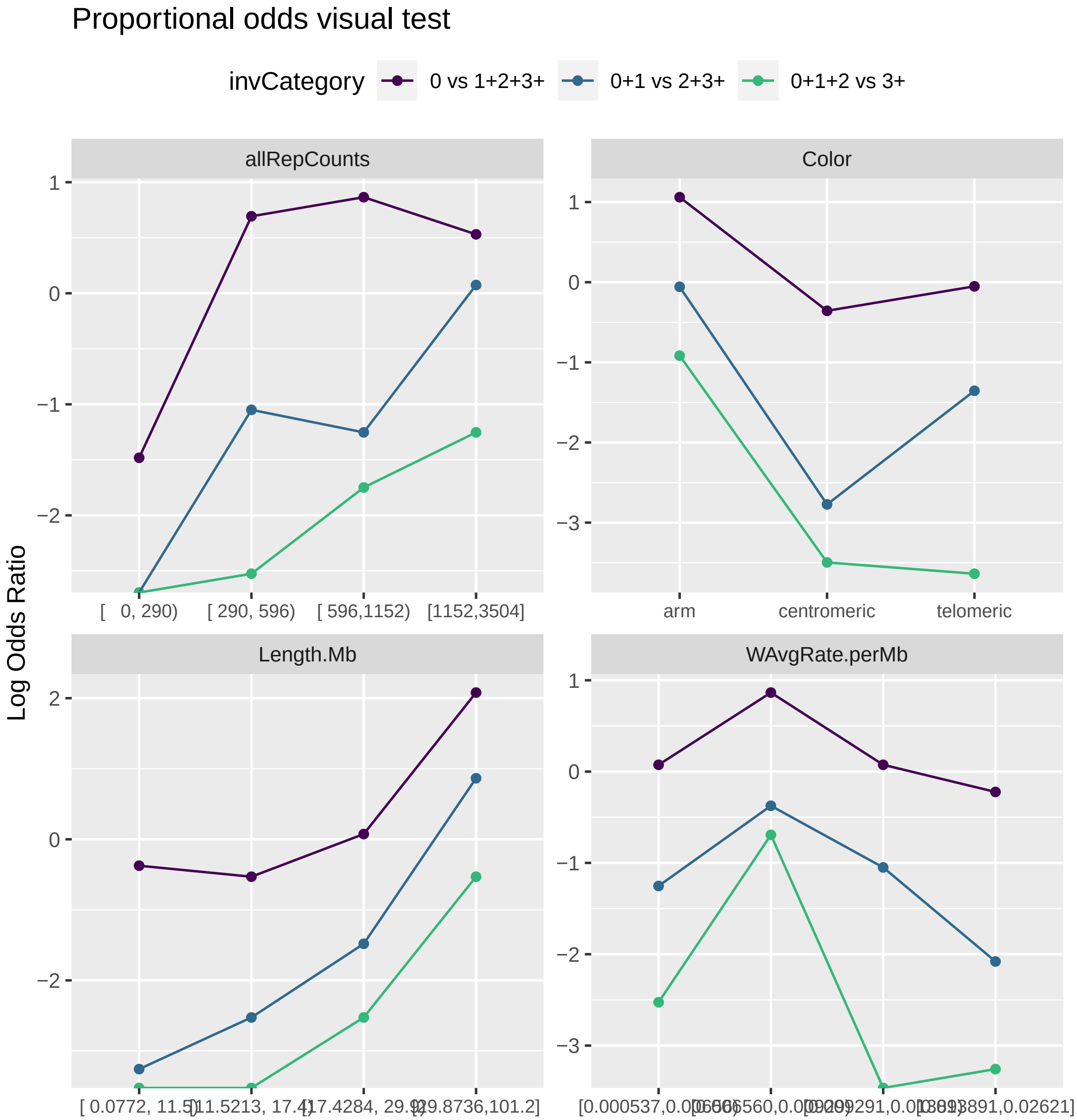
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          13.41   10  0.2
## Length.Mb        3.59    2   0.17
## allRepCounts     4.68    2   0.1
## Colorcentromeric 1.32    2   0.52
## Colortelomeric       3.18    2   0.2
## WAvgRate.perMb       2.29    2   0.32
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

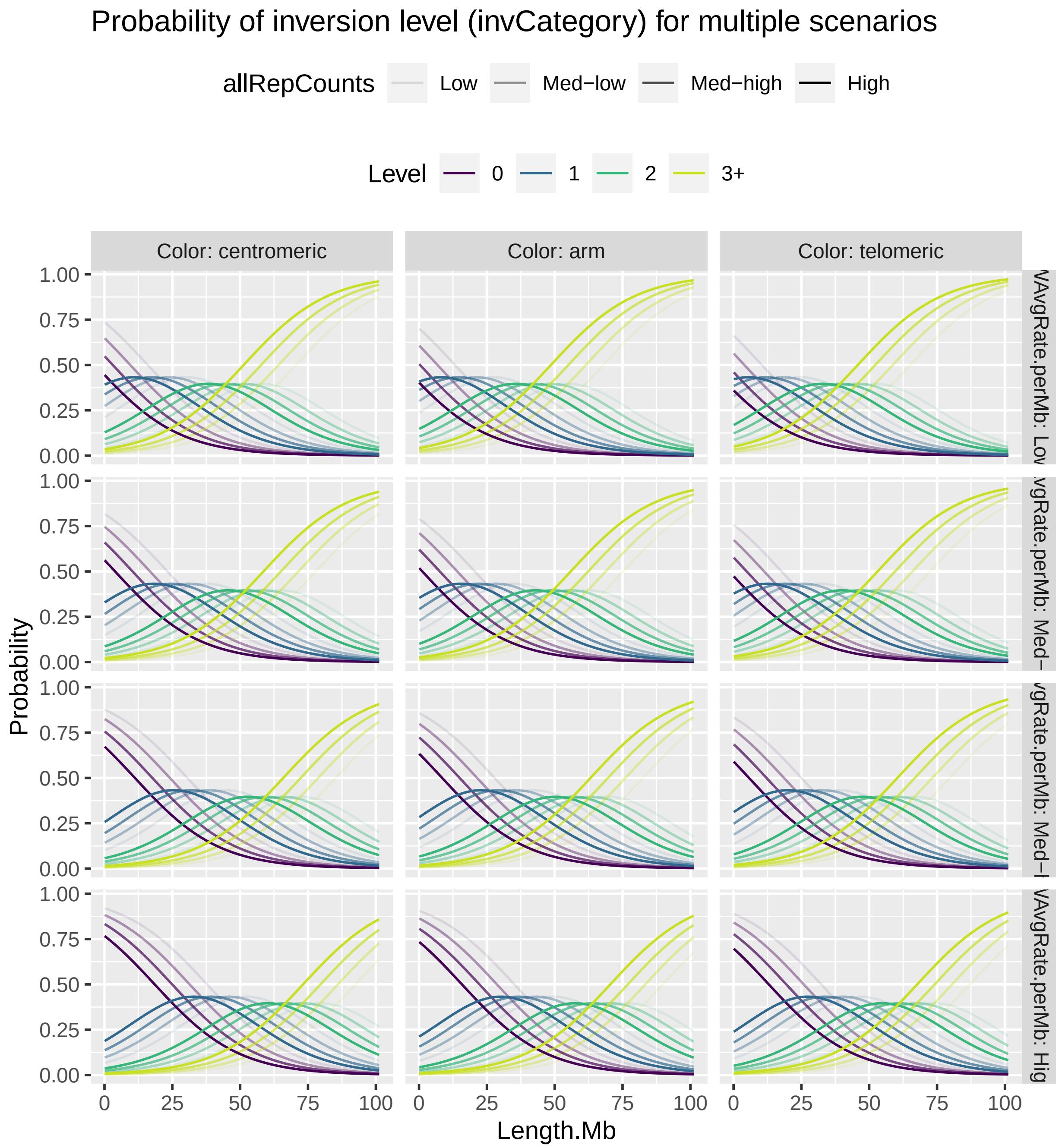


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Total inversions (NHCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                        Value Std. Error   t value
## Length.Mb          0.0861485  0.0164166    5.2476
## allRepCounts      -0.0002485  0.0003387   -0.7338
## Colorcentromeric   0.5947954  0.7347092    0.8096
## Colortelomeric     0.1488707  0.6177909    0.2410
## WAvgRate.perMb    -1.8225514  0.0128070 -142.3094
##
## Intercepts:
##      Value     Std. Error t value
## 0|1     2.4277    0.7981     3.0417
## 1|2     4.5520    0.9192     4.9523
## 2|3+    5.7942    1.0397     5.5732
##
## Residual Deviance: 172.2764
## AIC: 188.2764
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                         Value   Std. Error      t value    p value
## Length.Mb          0.0861484832 0.0164166341    5.2476338 0.00000015
## allRepCounts      -0.0002485214 0.0003386562   -0.7338456 0.46304285
## Colorcentromeric   0.5947954260 0.7347092391    0.8095657 0.41818981
## Colortelomeric     0.1488707219 0.6177909289    0.2409727 0.80957631
## WAvgRate.perMb    -1.8225514047 0.0128069632 -142.3094122 0.00000000
## 0|1                2.4277149369 0.7981497389    3.0416785 0.00235263
## 1|2                4.5519502700 0.9191660608    4.9522610 0.00000073
## 2|3+               5.7941831693 1.0396503917    5.5732035 0.00000003
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                          2.5 %       97.5 %
## Length.Mb          0.0500673141 0.1278096335
## allRepCounts      -0.0009522115 0.0003792915
## Colorcentromeric  -1.0986944207 2.4423927326
## Colortelomeric    -1.1552097209 1.5476442687
## WAvgRate.perMb              NA           NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                          2.5 %        97.5 %
## Length.Mb          0.0539724716  0.1183244948
## allRepCounts      -0.0009122753  0.0004152326
## Colorcentromeric  -0.8452082219  2.0347990738
## Colortelomeric    -1.0619772488  1.3597186925
```

```
## WAvgRate.perMb   -1.8476525913 -1.7974502181
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                  Odds Ratio      2.5%     97.5%
## Length.Mb         1.0899682 1.0513419  1.136337
## allRepCounts      0.9997515 0.9990482  1.000379
## Colorcentromeric  1.8126601 0.3333060 11.500526
## Colortelomeric    1.1605229 0.3149915  4.700384
## WAvgRate.perMb    0.1616129        NA        NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0899682 times more likely to increase in inversion amount category."

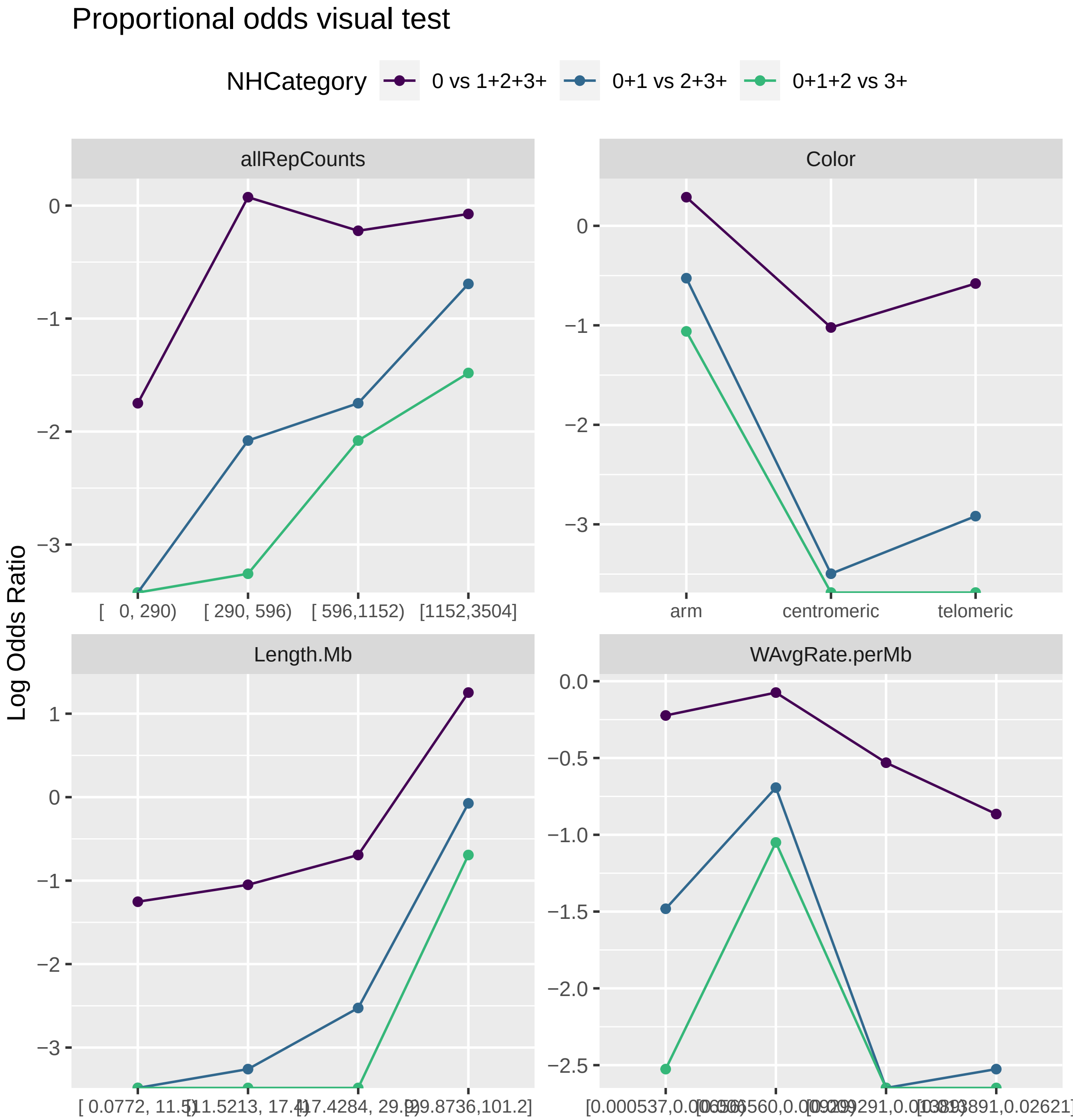
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          7.42    10  0.69
## Length.Mb        5.91    2   0.05
## allRepCounts     0.51    2   0.78
## Colorcentromeric 3.31    2   0.19
## Colortelomeric       0.02    2   0.99
## WAvgRate.perMb       7.3 2   0.03
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

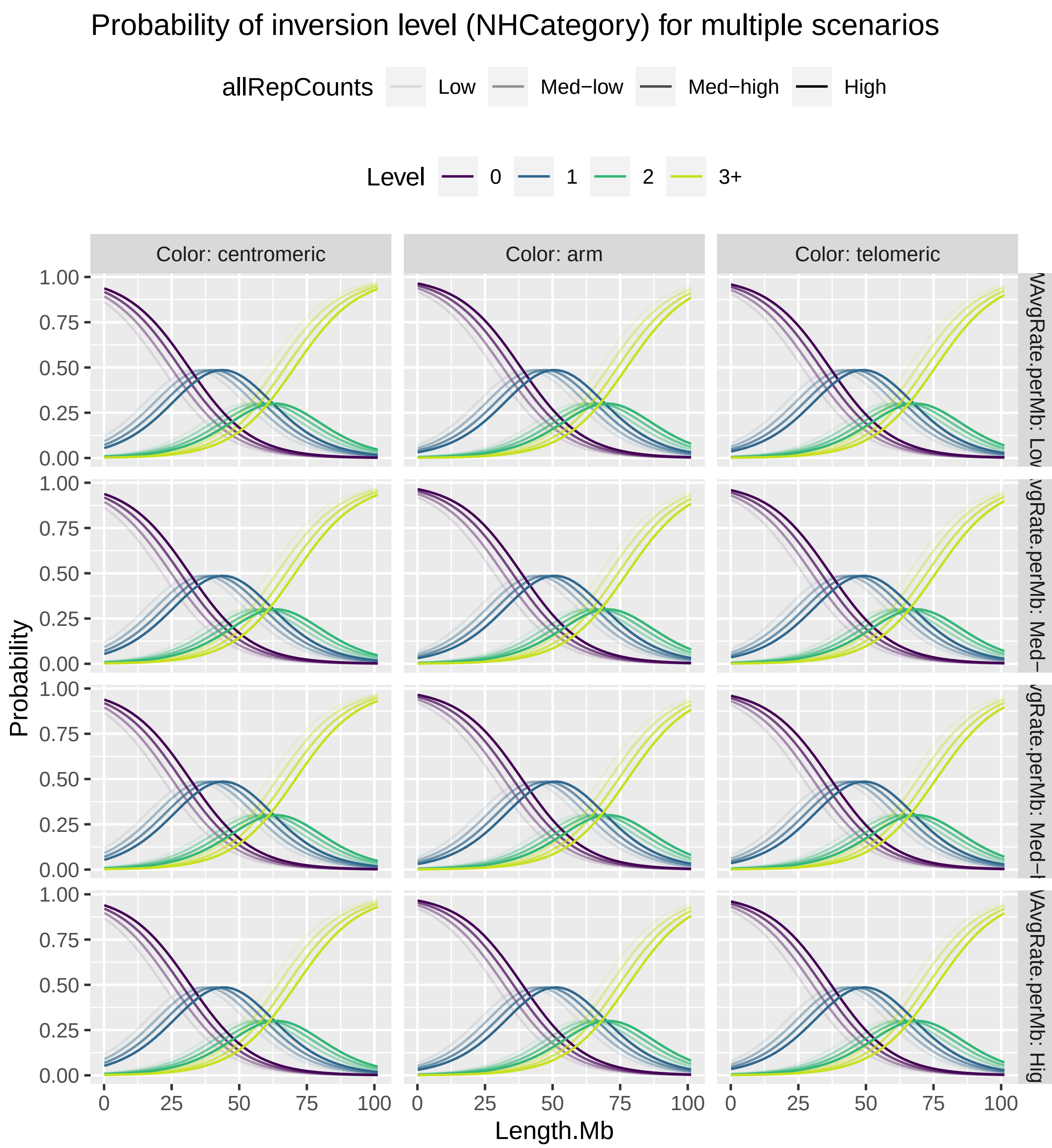


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Total inversions (NAHRCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error    t value
## Length.Mb          1.163e-03  0.0129147  9.004e-02
## allRepCounts       6.912e-04  0.0003706  1.865e+00
## Colorcentromeric -9.202e-01  0.7459849 -1.234e+00
## Colortelomeric     6.226e-01  0.6118582  1.017e+00
## WAvgRate.perMb   -1.120e+02  0.0101283 -1.106e+04
##
## Intercepts:
##     Value       Std. Error  t value
## 0|1      0.4281      0.7424      0.5766
## 1|2      2.1137      0.8263      2.5580
##
## Residual Deviance: 150.8857
## AIC: 164.8857
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                          Value   Std. Error       t value    p value
## Length.Mb          1.162857e-03 0.0129146512  9.004171e-02 0.92825407
## allRepCounts       6.912008e-04 0.0003705897  1.865138e+00 0.06216207
## Colorcentromeric -9.201954e-01 0.7459848525 -1.233531e+00 0.21737774
## Colortelomeric     6.225518e-01 0.6118582078  1.017477e+00 0.30892648
## WAvgRate.perMb   -1.119705e+02 0.0101283394 -1.105517e+04 0.00000000
## 0|1                4.280898e-01 0.7424062460  5.766248e-01 0.56419297
## 1|2                2.113749e+00 0.8263152782  2.558042e+00 0.01052634
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                         2.5 %      97.5 %
## Length.Mb        -2.819606e-02 0.032410324
## allRepCounts       3.421473e-05 0.001336844
## Colorcentromeric -2.635187e+00 0.850502401
## Colortelomeric   -7.170723e-01 2.000262576
## WAvgRate.perMb              NA          NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                         2.5 %        97.5 %
## Length.Mb        -2.414939e-02  2.647511e-02
## allRepCounts     -3.514156e-05  1.417543e-03
## Colorcentromeric -2.382299e+00  5.419081e-01
## Colortelomeric   -5.766683e-01  1.821772e+00
## WAvgRate.perMb   -1.119903e+02 -1.119506e+02
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                     Odds Ratio       2.5%    97.5%
## Length.Mb        1.001164e+00 0.97219774 1.032941
## allRepCounts     1.000691e+00 1.00003422 1.001338
## Colorcentromeric 3.984412e-01 0.07170558 2.340823
## Colortelomeric   1.863678e+00 0.48817943 7.390997
## WAvgRate.perMb   2.354157e-49         NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0011635 times more likely to increase in inversion amount category."

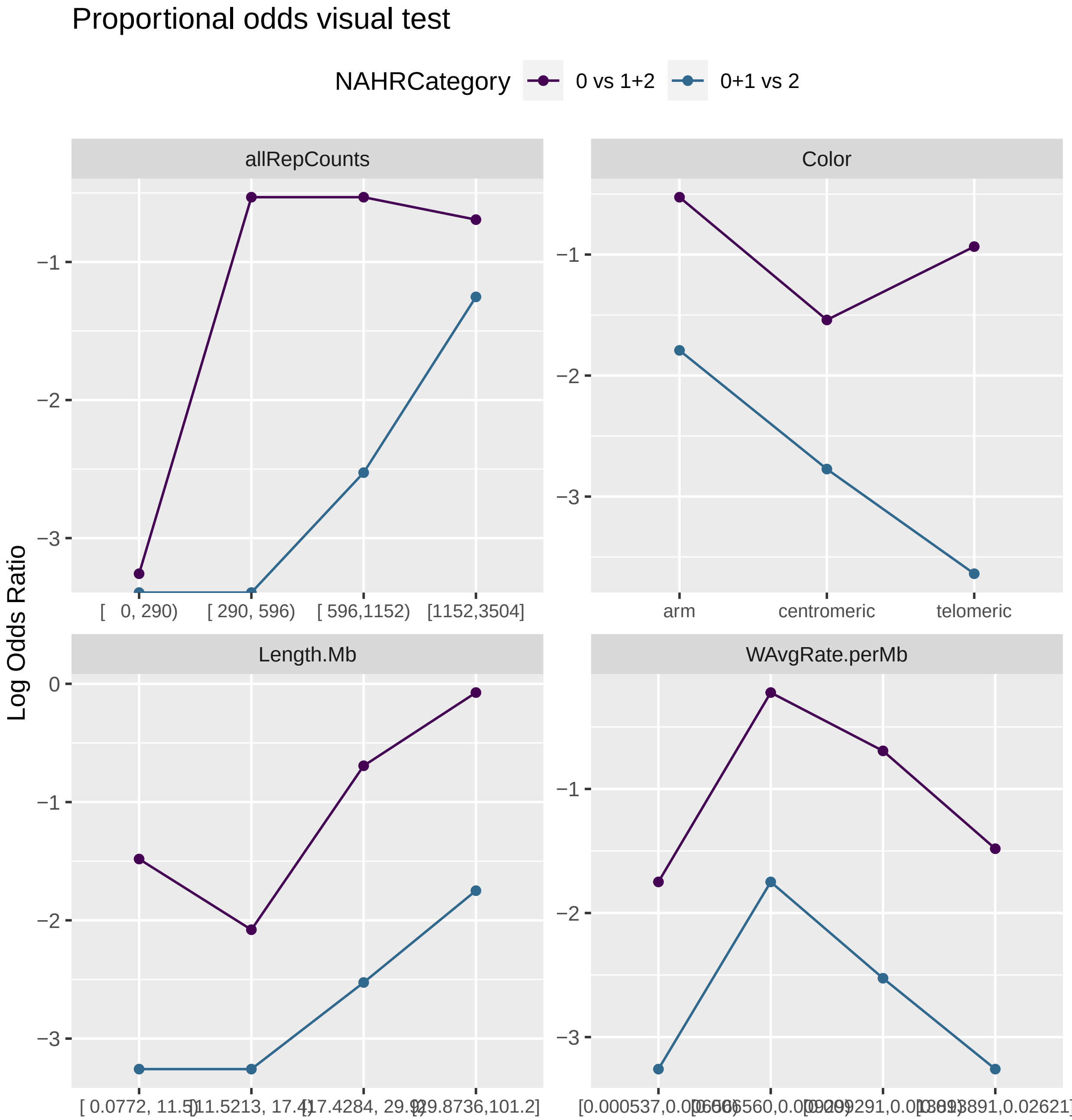
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          34.57   5   0
## Length.Mb        0.49    1   0.49
## allRepCounts     4.31    1   0.04
## Colorcentromeric 0.02    1   0.89
## Colortelomeric       2.35    1   0.13
## WAvgRate.perMb       0.46    1   0.5
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

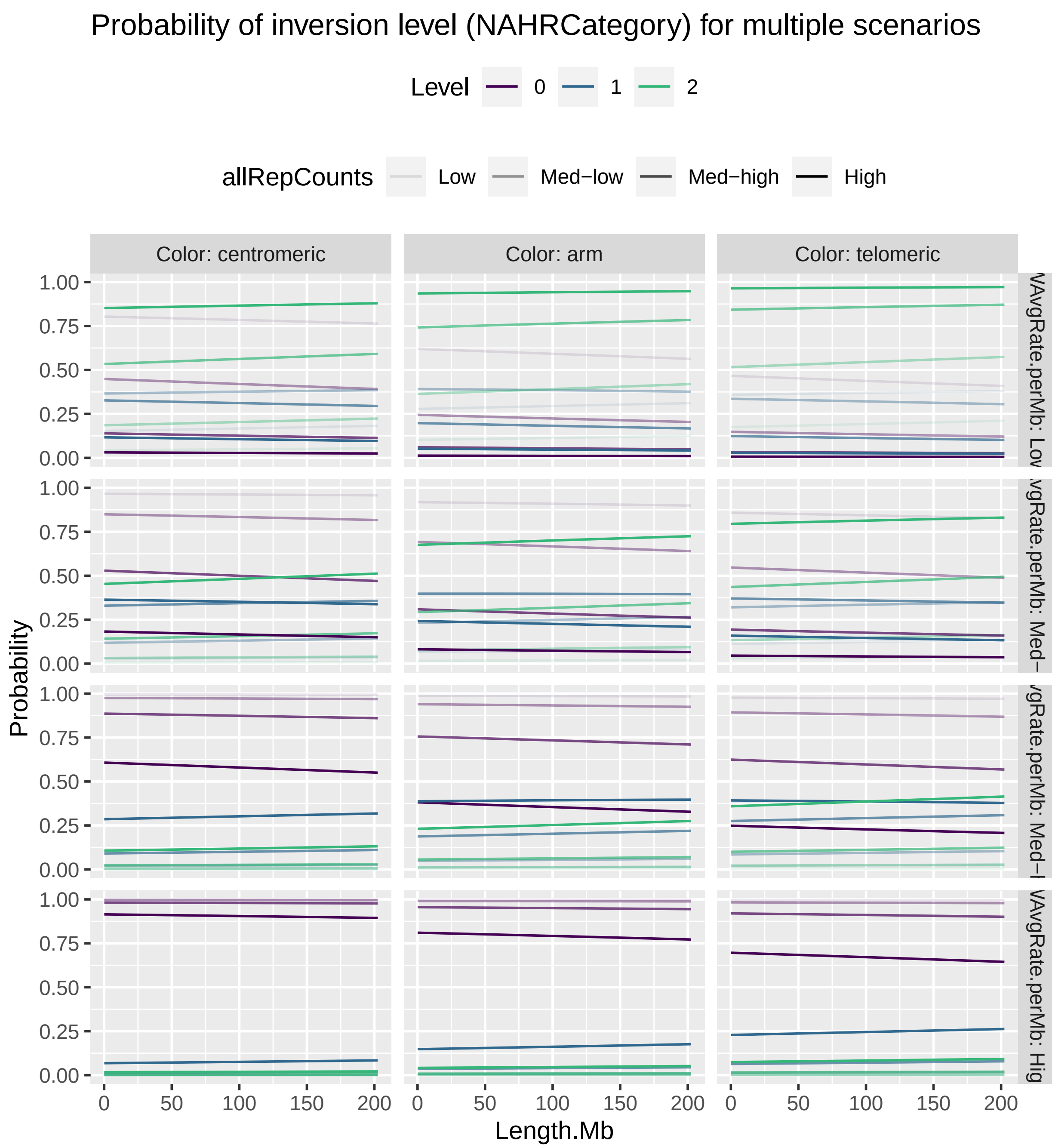


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables