Chromosomal aberrations and inversions in heterozygosis

Ruth Gómez Graciani

In this report, I analyze the realtionship between the amount of chromosomes in heterozygosis and amount of aneuploidies.

1 The data

If a recombination event takes place within an inversion region in heterozygosis, dicentric and acentric chromosomes will be formed. Later, the dicentric will breake at a random point between is two centromeres. In Bell et al. 2020, besides recombination information, they also provide chromosomal aberration counts by chromosome and individual (Table 1), in which the amount of inversions in heterozygosis could have a role. To simplify the analysis, I selected the columns corresponding to gains and losses of whole chromosomes, and gains and losses of chromosome arms. Figure 1 shows the relative values in aberrations per cell.

Table 1: Sample aneuploidy table

donor	chr	totalStructuralVariants	totalWholeAneuploidies	totalArmLevelStructuralVariants	totalLosses	totalWholeLosses	totalArmLosses	totalGains	totalWholeGains	totalArmGains	totalGainsOfOneCopy	totalGainsOfMoreThanOneCopy	totalWholeGainsOfOneCopy	totalArmGainsOfOneCopy	totalWholeGainsOfMoreThanOneCopy
all	overall	920	787	133	616	554	62	304	233	71	267	37	214	53	19
all	chr1	41	21	20	29	15	14	12	6	6	9	3	6	3	0
all	chr2	43	38	5	25	25	0	18	13	5	14	4	11	3	2
all	chr3	23	16	7	12	11	1	11	5	6	9	2	5	4	0
all	chr4	30	20	10	19	17	2	11	3	8	9	2	3	6	0
all	chr5	17	14	3	10	9	1	7	5	2	7	0	5	2	0

The quantity of cM affected by the presence of an inversion in heterzygousis would be the determinant factor for a possible increase of chromosomal aberrations. Genetic size measurements for each inversion and individual are obtained by transforming the normalized recombination rates in cM/Mb to cM and making the mean of the available measurements. Figure 2 shows value distributions and correlation between physical and genetic size for inversion genetic sizes calculated with the mean of STD, HOM and HET individuals. Since we expect recombination rates to be affected in heterozygous individuals, I consider the genetic size of each inversion to be the mean genetic sizes of the region in homozygous individuals. Figure 3 shows the distribution of inversions, in counts and genetic sizes, across the different chromosomes, which gives an idea of the probability of a chromosome to be affected by recombination events taking place in an inversion region in heterozygosis.

2 Variables to test

In this analysis, the dependent variable is count of aberrations per cell (as mentioned before, we will look at different types of aberrations), and the independent variable to test for inversions in heterozygosis effect will be the sum of cM affected by heterozygous inversions in each chromosome. In addition, crossovers are suggested to protect against chromosomal aberrations:

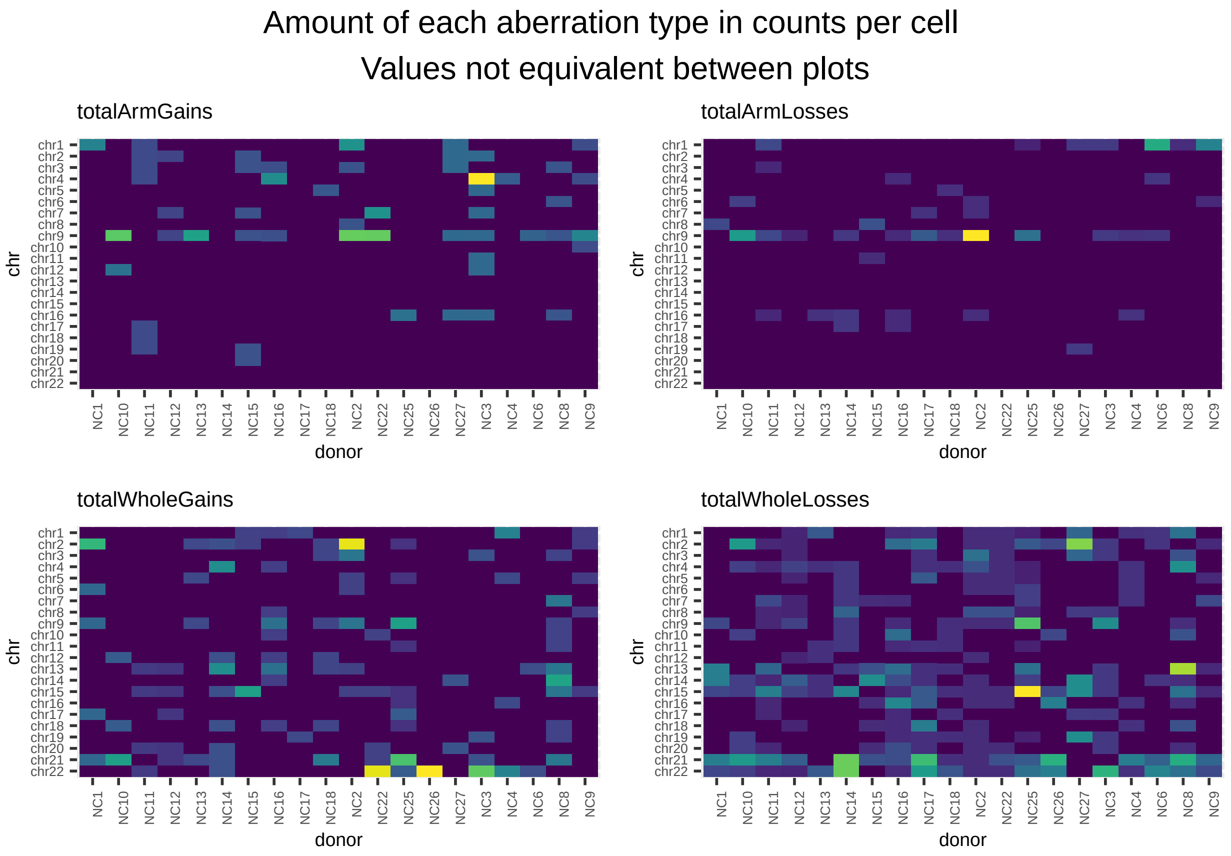


Figure 1: Amount of different aberration types in counts per cell by individual and chromosome. Colors are relative, not equivalent between plots.

> We found that chromosome gains originating in meiosis I—when recombination occurs—had 36% fewer total crossovers than matched, well-segregated chromosomes did (Supplementary Methods), suggesting that crossovers protected against meiosis I nondisjunction of the chromosomes on which they occurred (Extended Data Fig. 9b and Supplementary Notes). No similar relationship was observed for meiosis II gains (although the simulated control distribution for meiosis II is inherently less accurate; Supplemen- tary Notes) or at other levels of aggregation (Extended Data Fig. 9b–d and Supplementary Notes).
>
> — Bell et al. 2020

So we should take into consideration the amount of crossovers, which I could measure as crossovers per cell in each chromosome (crossovers per chromosome) or as crossovers per cell in general.

> Another study can be total aberrations per individual with crossovers per cell and total cM affected.
>
> — Side note

A multiple regression is suitable to test for inversions in heterozygosis effect while controlling the recombination rate effect as well:

> Use multiple regression when you have a more than two measurement variables, one is the dependent variable and the rest are independent variables. You can use it to predict values of the dependent variable, or if you're careful, you can use it for suggestions about which independent variables have a major effect on the dependent variable.
>
> — Handbook of biological statistics

However, first we have to know the distribution of the variables and the relationships between them:

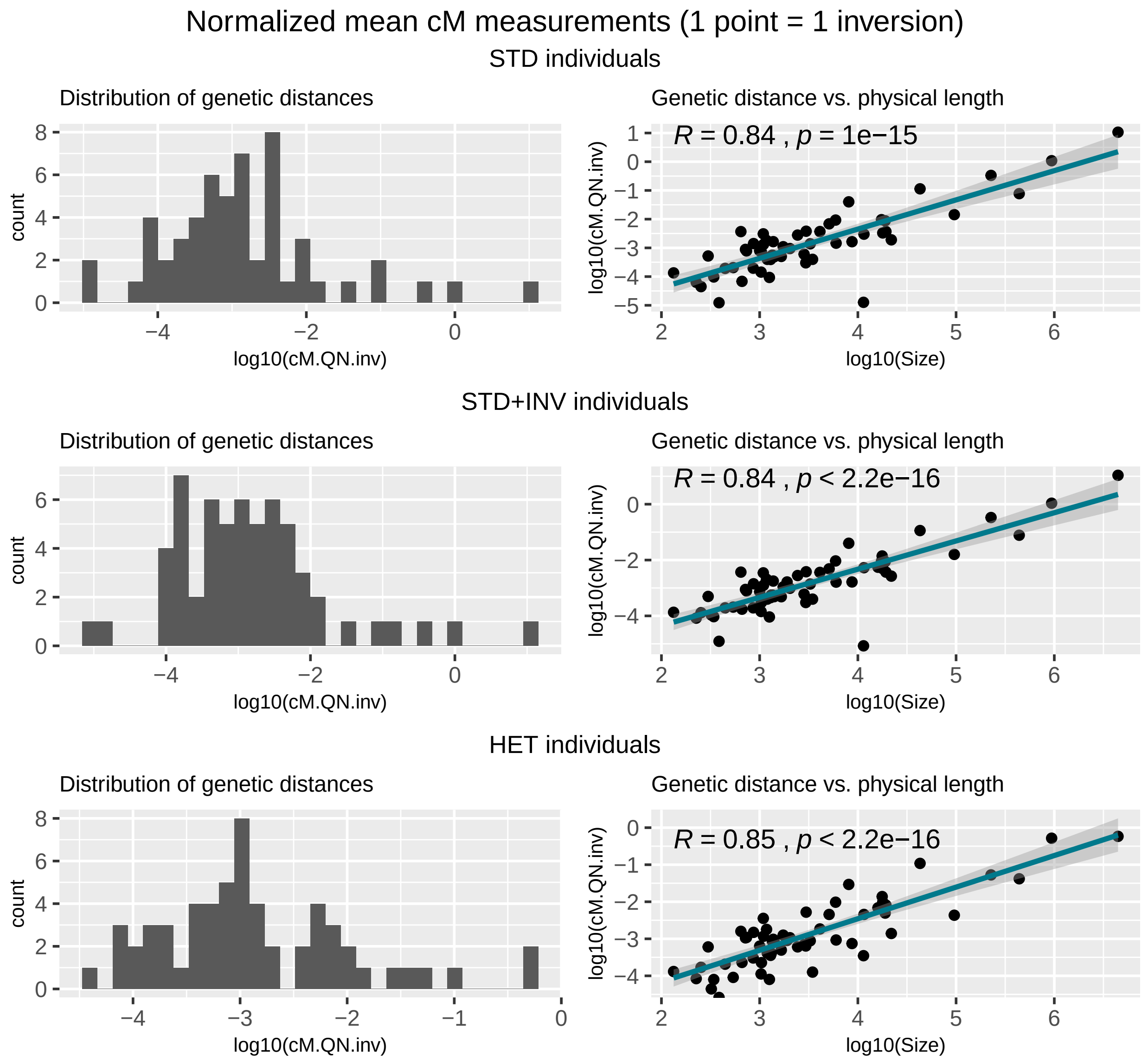


Figure 2: Value distributions and correlation between physical and genetic size for inversion genetic sizes calculated with the mean of STD, HOM and HET individuals

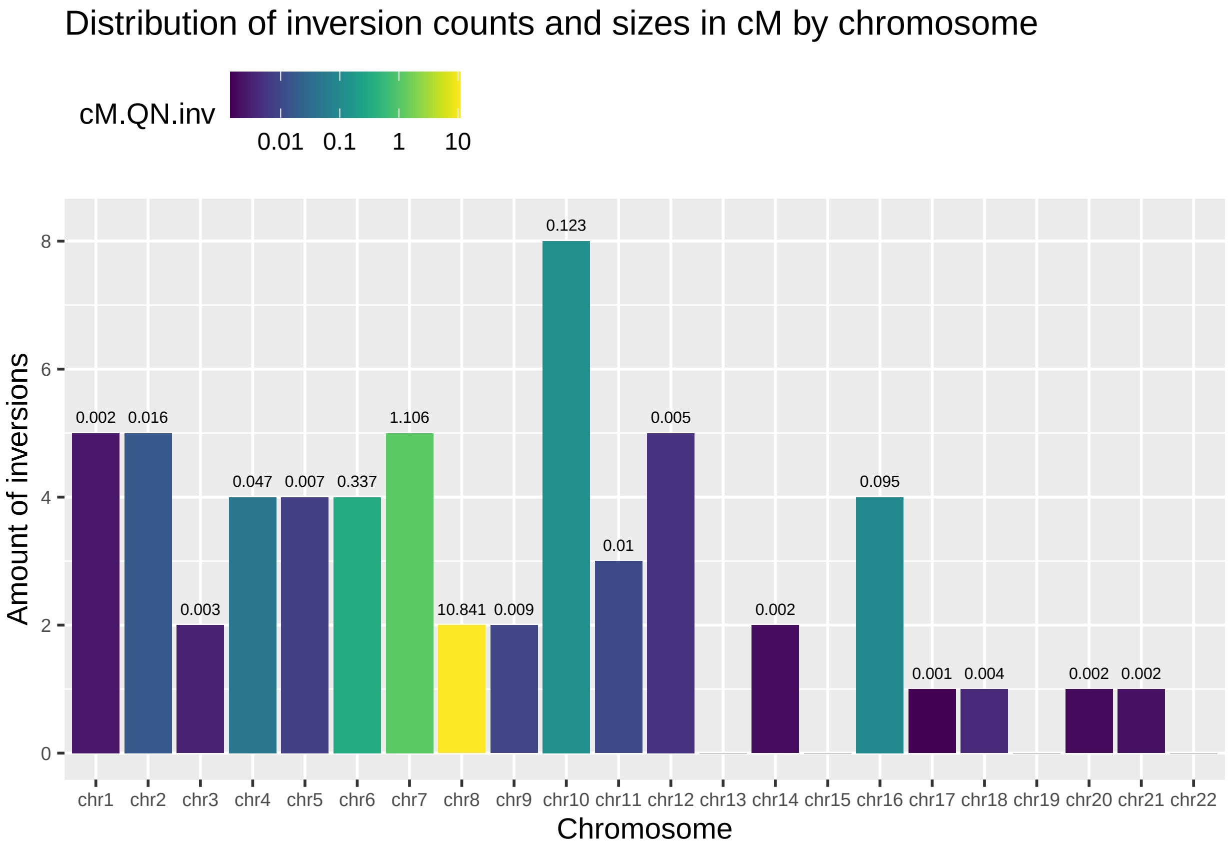


Figure 3: Bar plots represent amount of inversions in each chromosome, colored by the sum of inversions sizes in cM, which is also indicated by bar colors. This sum of inversion genetic sizes corresponds to the maximum cM affected, given that an individual in heterozygous for all the detectable inversions.

> Whenever you have a dataset with multiple numeric variables, it is a good idea to look at the correlations among these variables. One reason is that if you have a dependent variable, you can easily see which independent variables correlate with that dependent variable. A second reason is that if you will be constructing a multiple regression model, adding an independent variable that is strongly correlated with an independent variable already in the model is unlikely to improve the model much, and you may have good reason to chose one variable over another.
>
> Finally, it is worthwhile to look at the distribution of the numeric variables. If the distributions differ greatly, using Kendall or Spearman correlations may be more appropriate. Also, if independent variables differ in distribution from the dependent variable, the independent variables may need to be transformed.
>
> — Handbook of biological statistics

2.1 Variable distributions

Dependent variables (counts per cell of different types of aberrations) seem to have a distribution that turns more or less normal with a logarithmic transformation, as shown in figure 4.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1116 rows containing non-finite values (stat_bin).
```

Genetic distance in heterozygosis seems to work better with a logarithmic transformation as well (Figure 5).

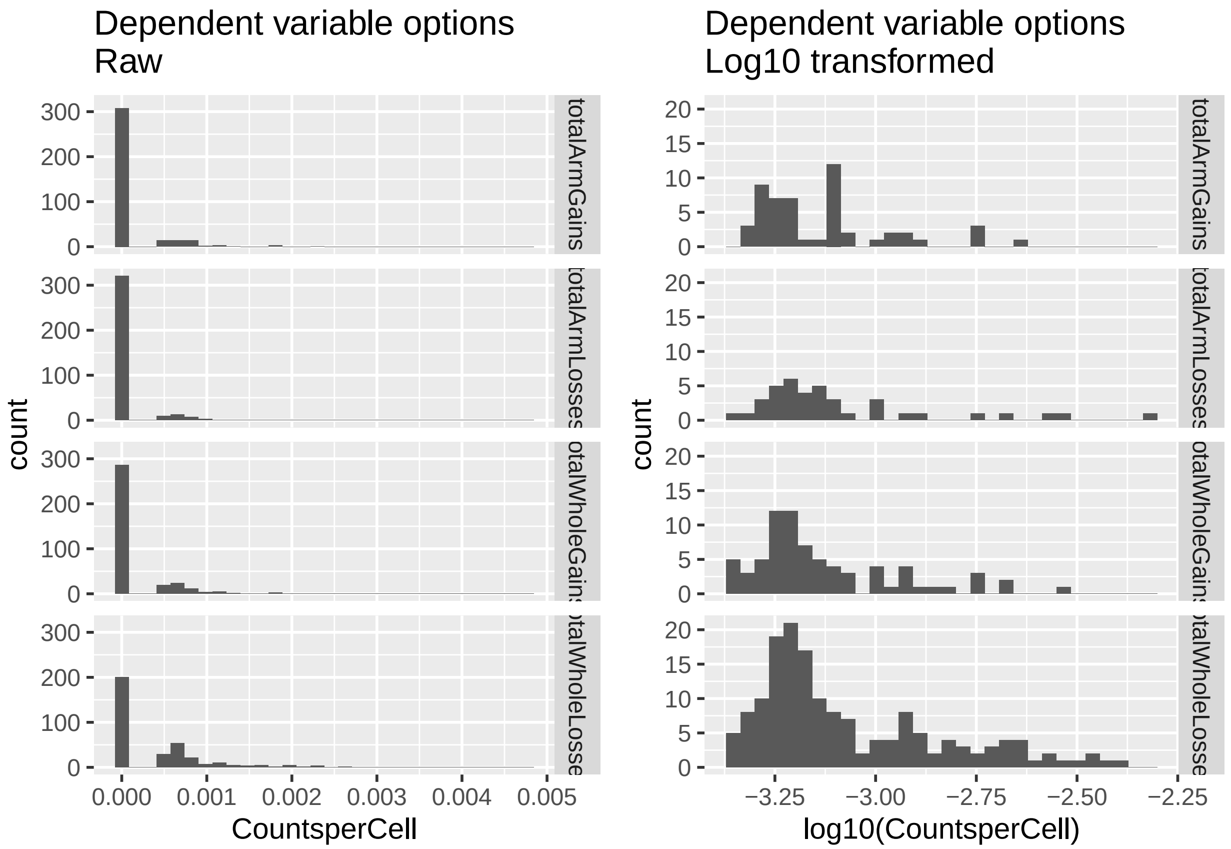


Figure 4: Dependent variables distribution, raw and log10 transformed

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 150 rows containing non-finite values (stat_bin).
```

Finally, variables related to global crossover distribution and per-chromosome crossover distribution are more or less normal already (Figure 6).

2.2 Variable correlations - raw data

The results, however, between our two main variables we want to look at, are better with untransformed data,

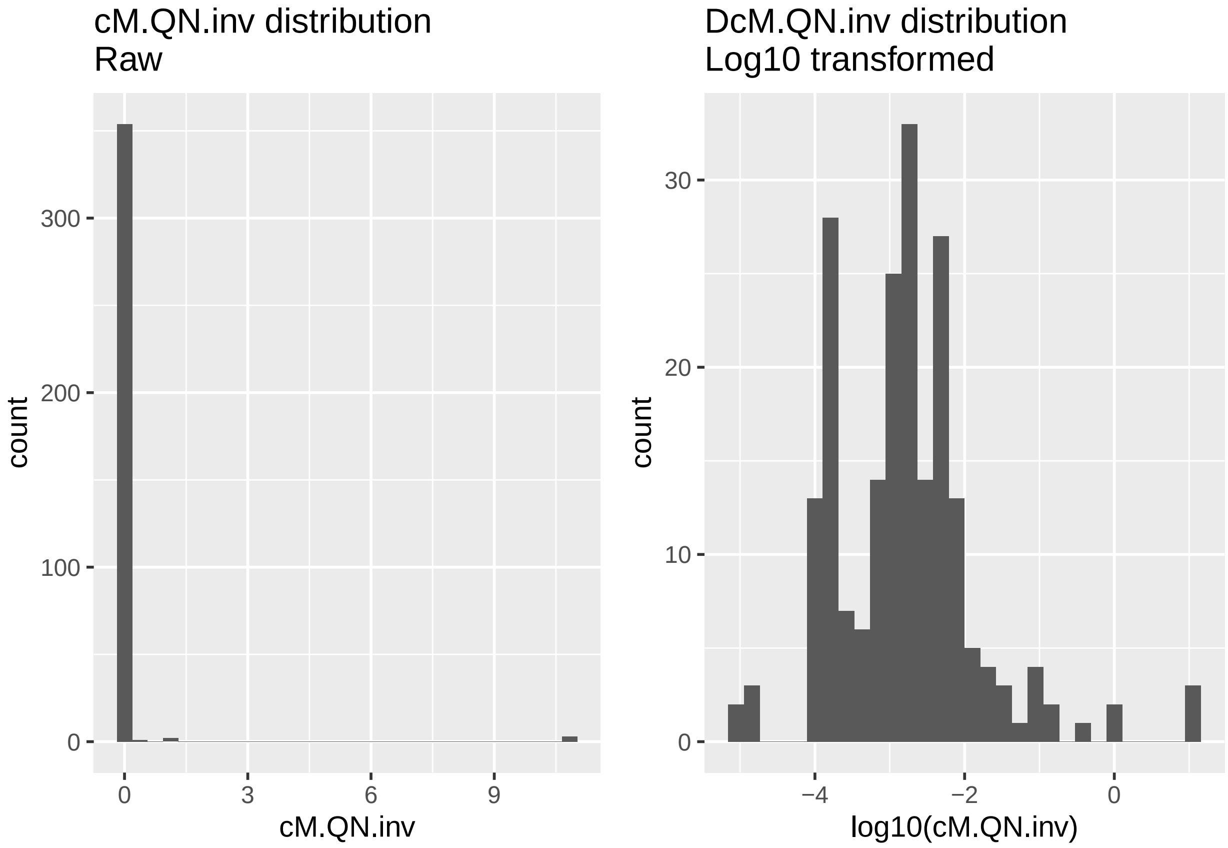


Figure 5: Dependent variables distribution, raw and log10 transformed

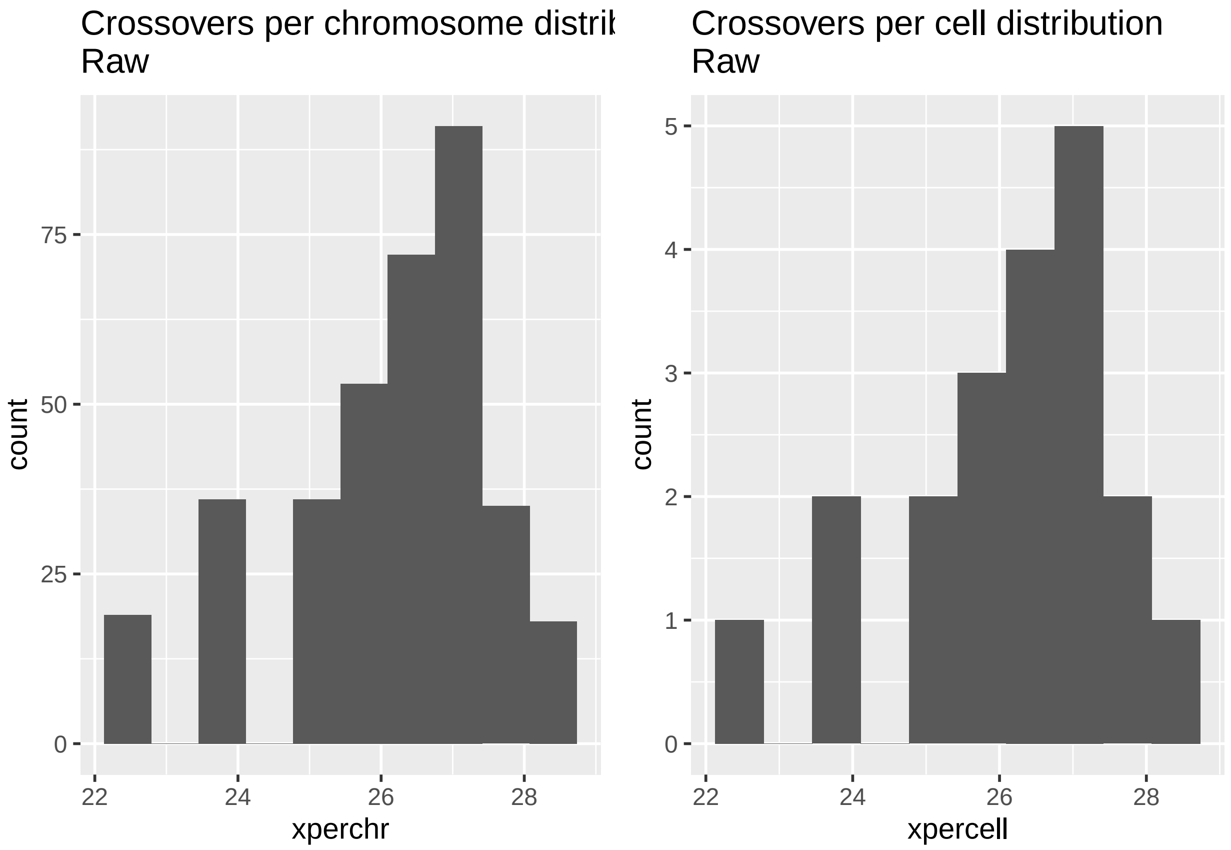
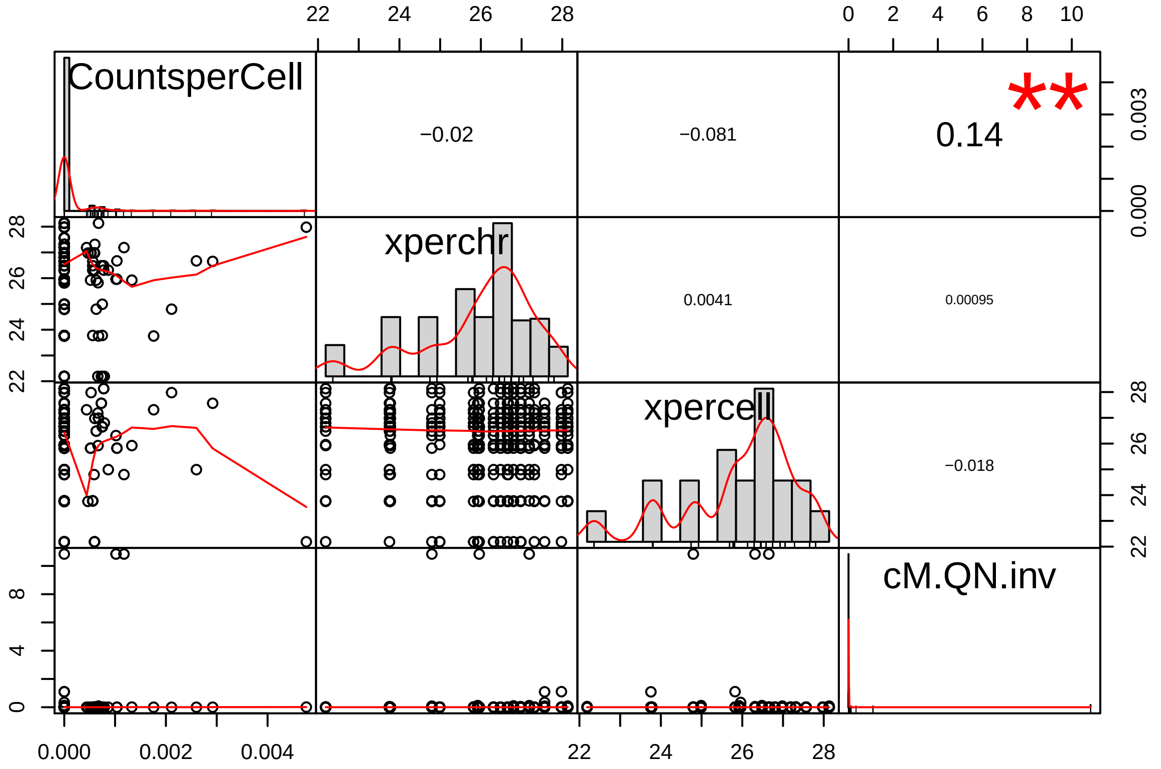


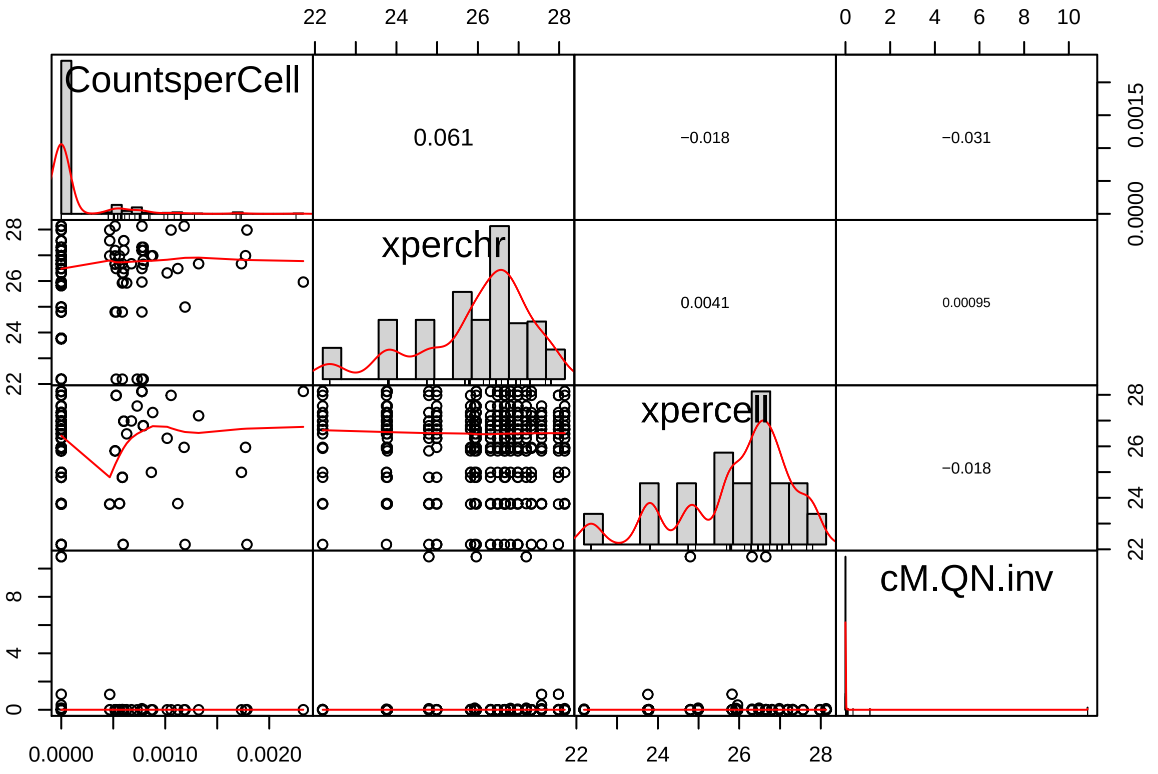
Figure 6: Dependent variables distribution, raw and log10 transformed

2.2.1 totalArmLosses



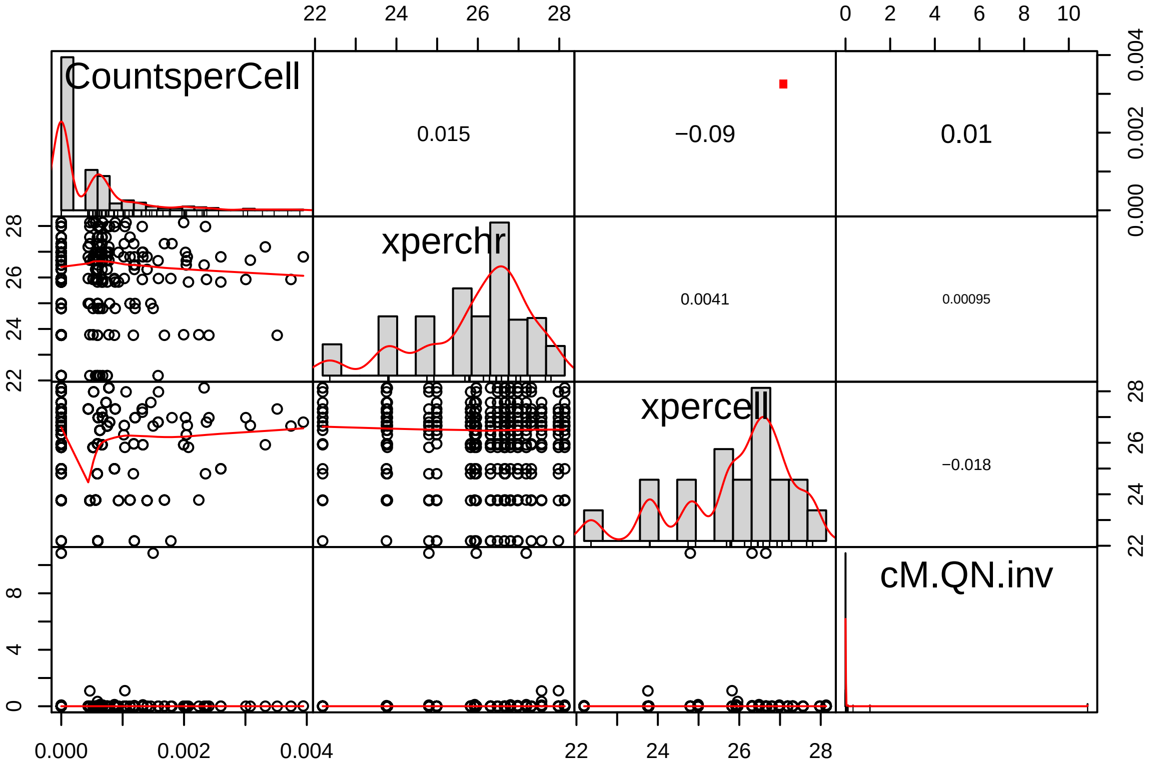
```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##              CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell         1.00   -0.02    -0.08      0.14
## xperchr              -0.02    1.00     0.00      0.00
## xpercell             -0.08    0.00     1.00     -0.02
## cM.QN.inv             0.14    0.00    -0.02      1.00
## Sample Size
## [1] 360
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##              CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell         0.00    0.71     0.13      0.01
## xperchr               0.71    0.00     0.94      0.99
## xpercell              0.13    0.94     0.00      0.74
## cM.QN.inv             0.01    0.99     0.74      0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.2.2 totalArmGains



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##              CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell         1.00    0.06    -0.02     -0.03
## xperchr               0.06    1.00     0.00      0.00
## xpercell             -0.02    0.00     1.00     -0.02
## cM.QN.inv            -0.03    0.00    -0.02      1.00
## Sample Size
## [1] 360
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##               CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell          0.00    0.25     0.73      0.55
## xperchr                0.25    0.00     0.94      0.99
## xpercell               0.73    0.94     0.00      0.74
## cM.QN.inv              0.55    0.99     0.74      0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

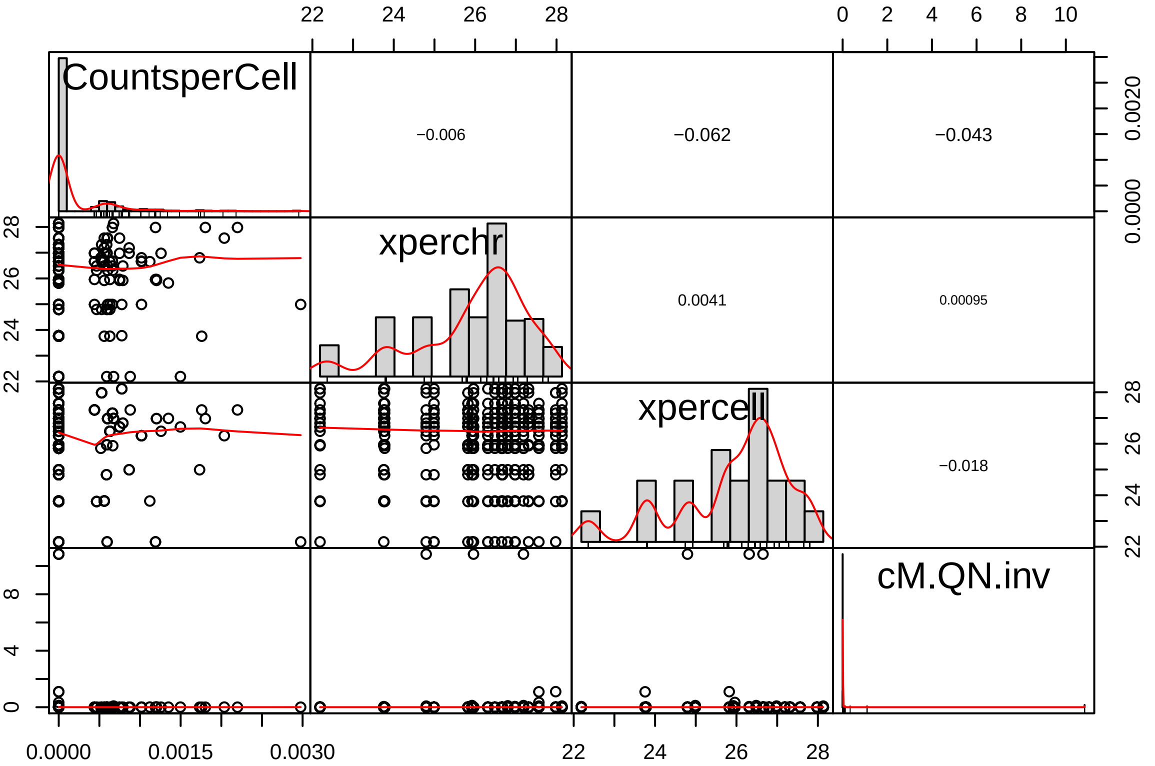
2.2.3 totalWholeLosses



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##               CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell          1.00    0.02    -0.09      0.01
## xperchr                0.02    1.00     0.00      0.00
## xpercell              -0.09    0.00     1.00     -0.02
## cM.QN.inv              0.01    0.00    -0.02      1.00
## Sample Size
## [1] 360
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##               CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell          0.00    0.78     0.09      0.84
## xperchr                0.78    0.00     0.94      0.99
## xpercell               0.09    0.94     0.00      0.74
## cM.QN.inv              0.84    0.99     0.74      0.00
```

```
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

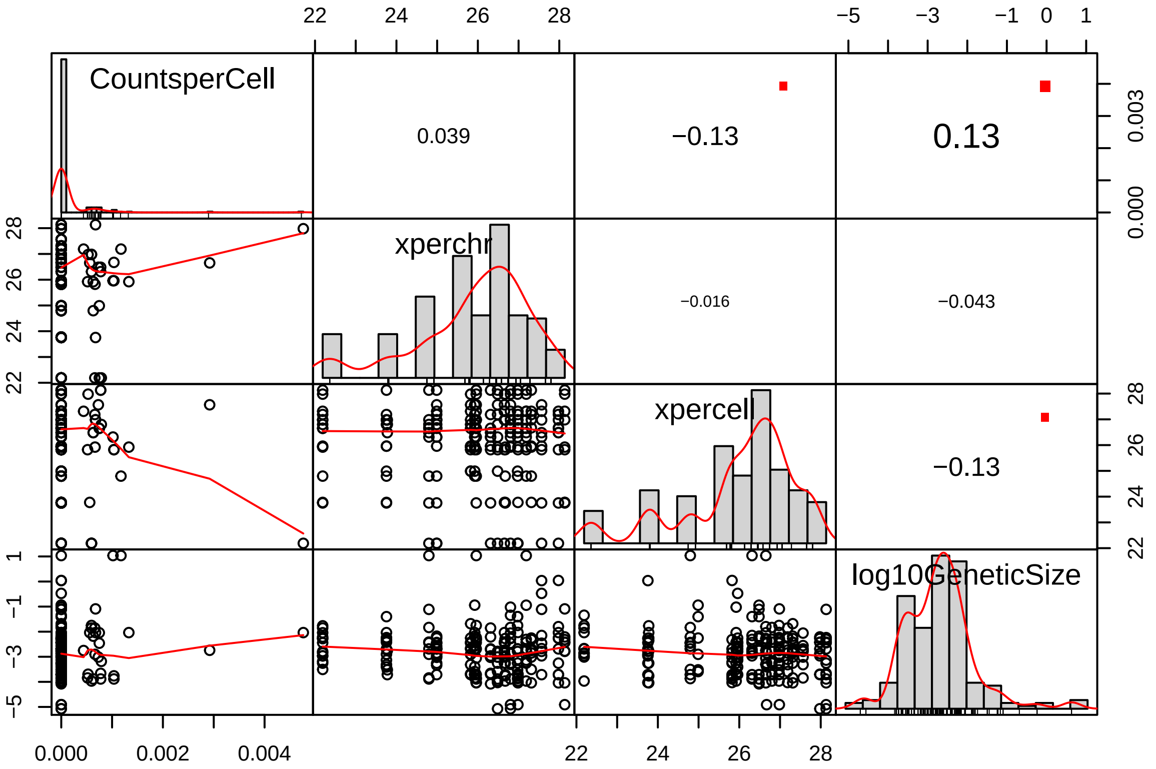
2.2.4 totalWholeGains



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##               CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell          1.00   -0.01    -0.06     -0.04
## xperchr               -0.01    1.00     0.00      0.00
## xpercell              -0.06    0.00     1.00     -0.02
## cM.QN.inv             -0.04    0.00    -0.02      1.00
## Sample Size
## [1] 360
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##               CountsperCell xperchr xpercell cM.QN.inv
## CountsperCell          0.00    0.91     0.24      0.42
## xperchr                0.91    0.00     0.94      0.99
## xpercell               0.24    0.94     0.00      0.74
## cM.QN.inv              0.42    0.99     0.74      0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

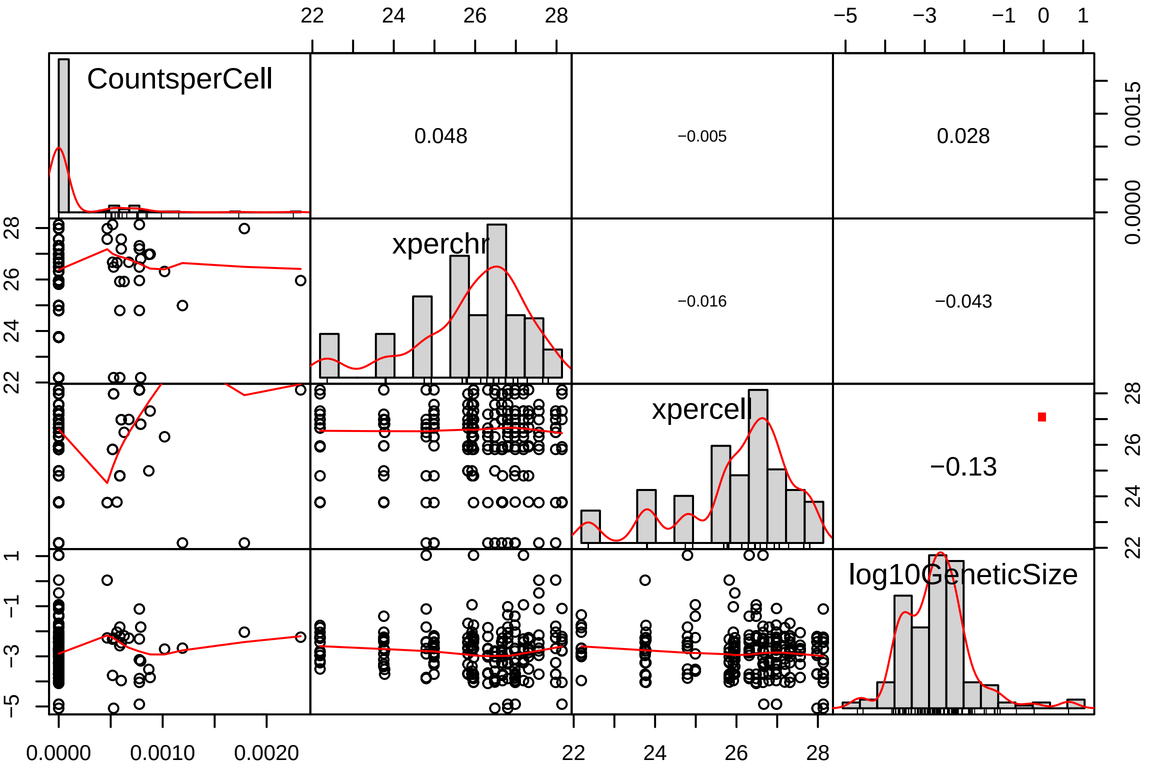
2.3 Variable correlations - partially transformed data

2.3.1 totalArmLosses



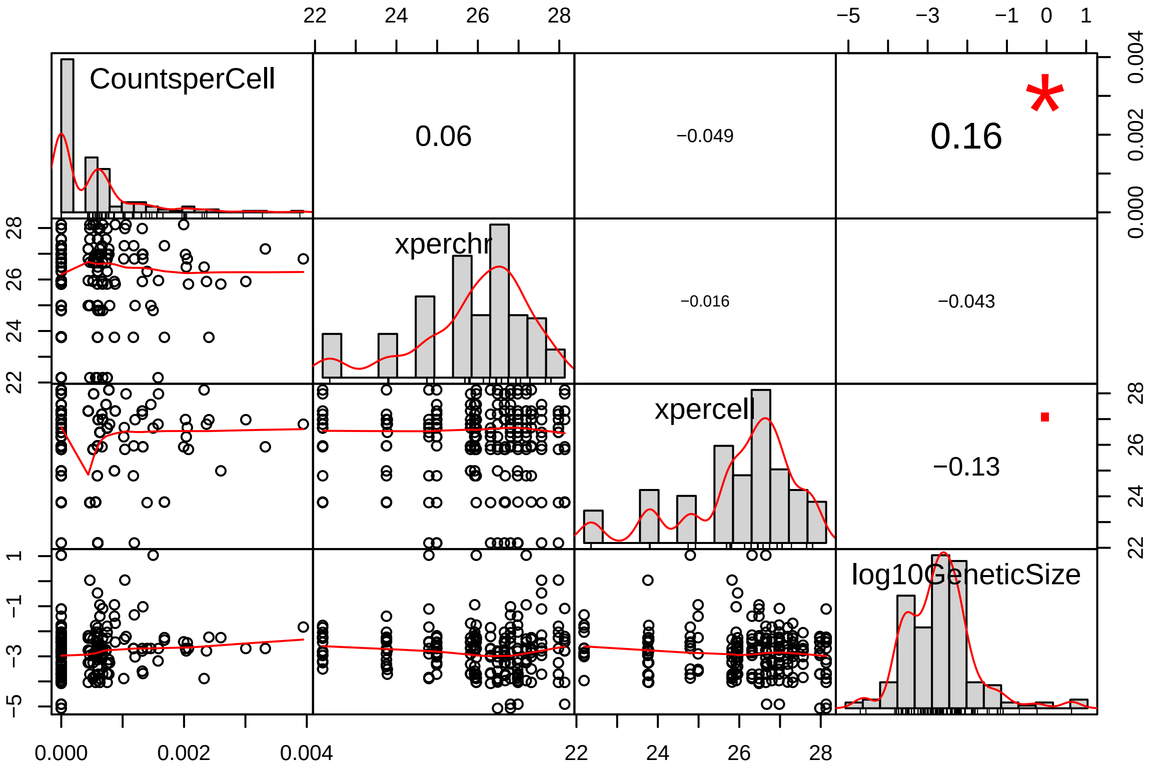
```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                    CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell               1.00    0.04    -0.13             0.13
## xperchr                     0.04    1.00    -0.02            -0.04
## xpercell                   -0.13   -0.02     1.00            -0.13
## log10GeneticSize            0.13   -0.04    -0.13             1.00
## Sample Size
## [1] 210
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                    CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell               0.00    0.57     0.06             0.07
## xperchr                     0.57    0.00     0.81             0.54
## xpercell                    0.06    0.81     0.00             0.07
## log10GeneticSize            0.07    0.54     0.07             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.3.2 totalArmGains



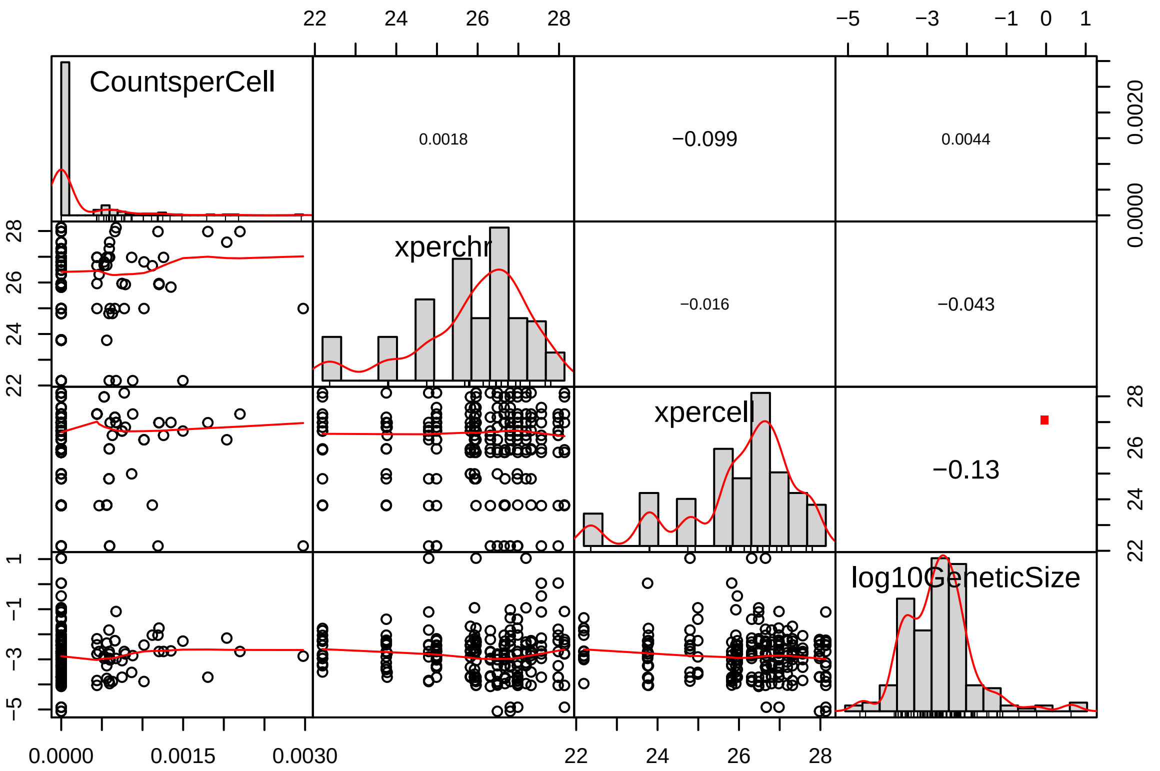
```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell             1.00    0.05    -0.01             0.03
## xperchr                   0.05    1.00    -0.02            -0.04
## xpercell                 -0.01   -0.02     1.00            -0.13
## log10GeneticSize          0.03   -0.04    -0.13             1.00
## Sample Size
## [1] 210
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell             0.00    0.49     0.94             0.69
## xperchr                   0.49    0.00     0.81             0.54
## xpercell                  0.94    0.81     0.00             0.07
## log10GeneticSize          0.69    0.54     0.07             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.3.3 totalWholeLosses



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell             1.00    0.06    -0.05             0.16
## xperchr                   0.06    1.00    -0.02            -0.04
## xpercell                 -0.05   -0.02     1.00            -0.13
## log10GeneticSize          0.16   -0.04    -0.13             1.00
## Sample Size
## [1] 210
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell             0.00    0.39     0.48             0.02
## xperchr                   0.39    0.00     0.81             0.54
## xpercell                  0.48    0.81     0.00             0.07
## log10GeneticSize          0.02    0.54     0.07             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.3.4 totalWholeGains

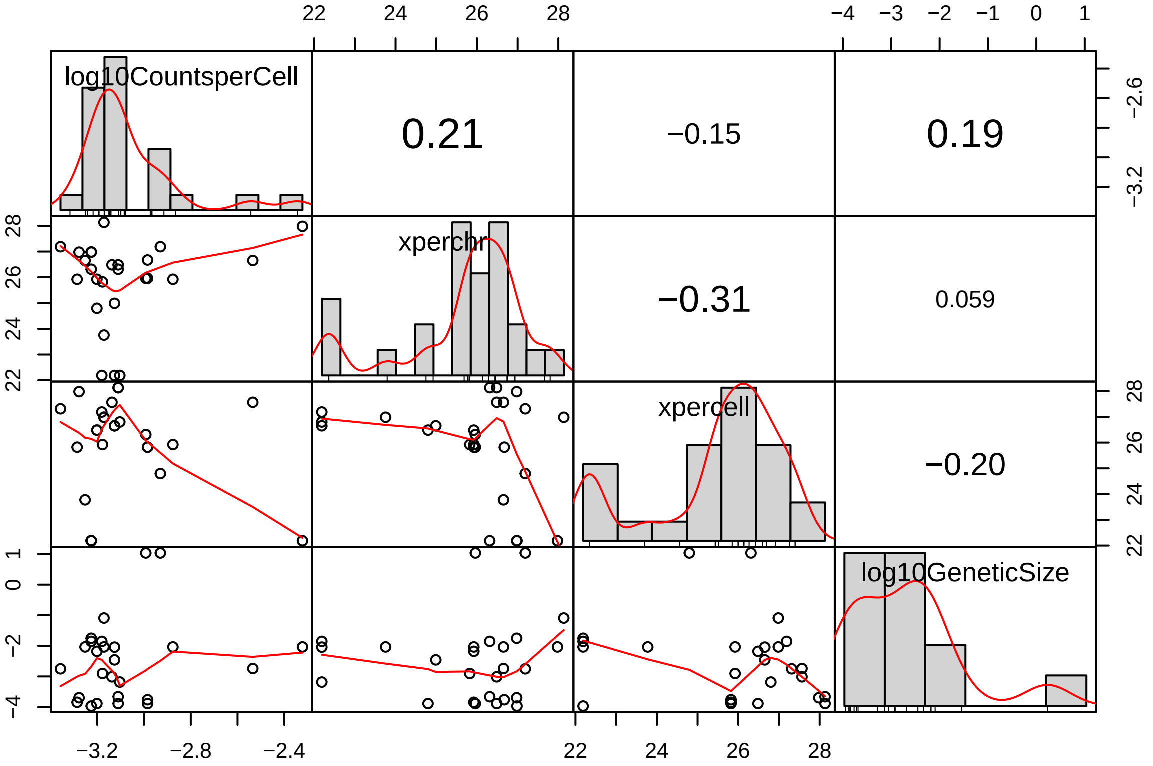


```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell              1.0    0.00    -0.10             0.00
## xperchr                    0.0    1.00    -0.02            -0.04
## xpercell                  -0.1   -0.02     1.00            -0.13
## log10GeneticSize           0.0   -0.04    -0.13             1.00
## Sample Size
## [1] 210
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                  CountsperCell xperchr xpercell log10GeneticSize
## CountsperCell             0.00    0.98     0.15             0.95
## xperchr                   0.98    0.00     0.81             0.54
## xpercell                  0.15    0.81     0.00             0.07
## log10GeneticSize          0.95    0.54     0.07             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.4 Variable correlations - transformed data

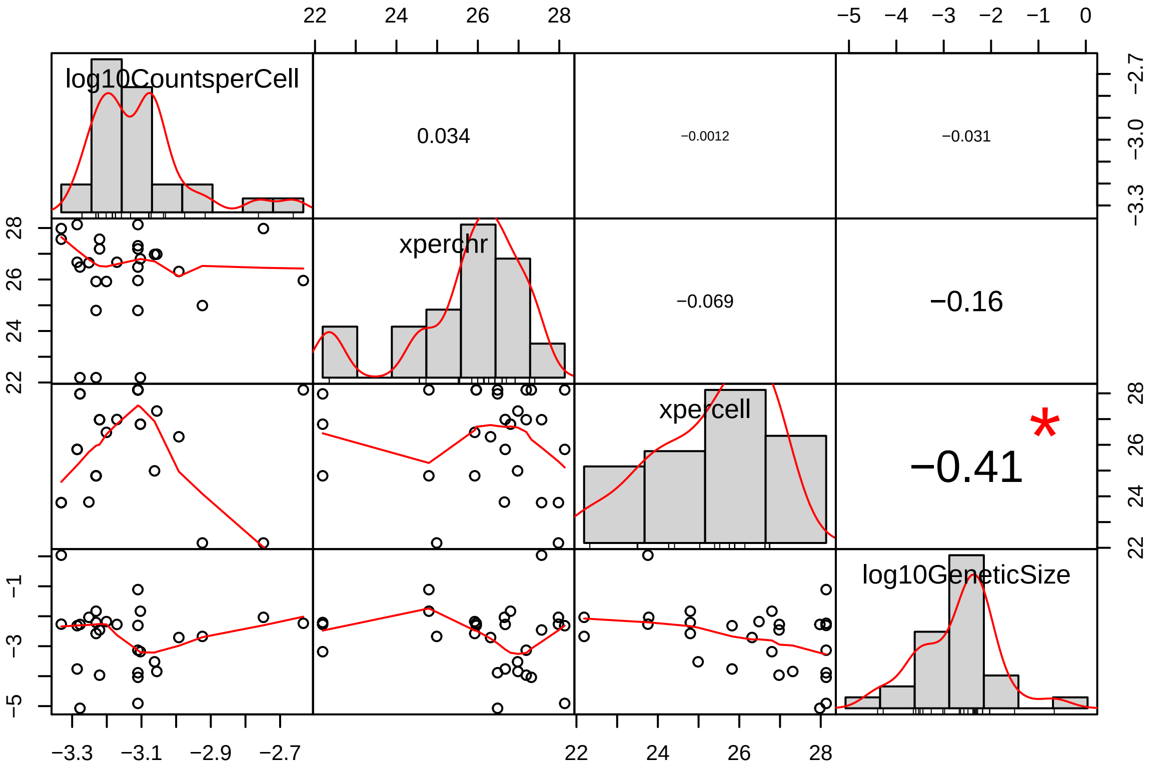
After doing a log10 transformation to the dependent variables and the genetic distance in heterozygosis (and removing points with -Inf values, which is required to perform the following analyses), the next step is to know the correlations among variables. With this information I will make sure that independent variables are actually independent, and select which of the methods to account for interindividual recombination rate variability is more useful.

2.4.1 totalArmLosses



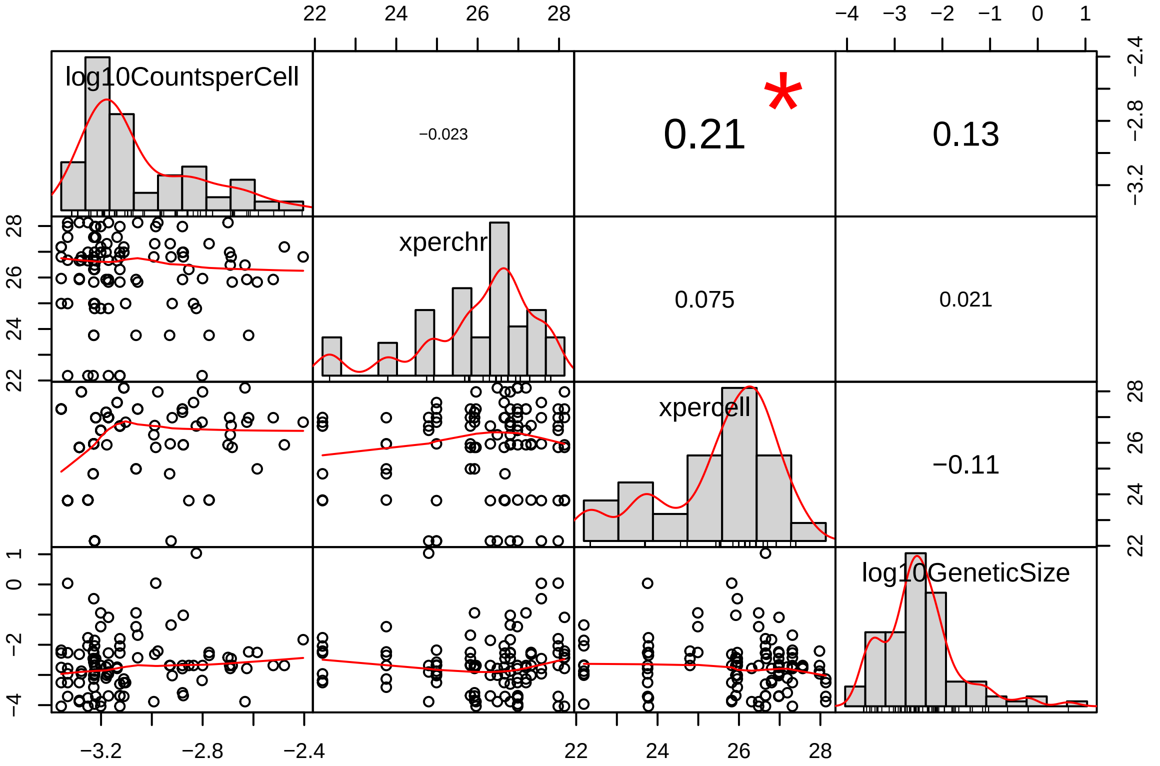
```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                      log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell                 1.00    0.21    -0.15             0.19
## xperchr                            0.21    1.00    -0.31             0.06
## xpercell                          -0.15   -0.31     1.00            -0.20
## log10GeneticSize                   0.19    0.06    -0.20             1.00
## Sample Size
## [1] 26
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                      log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell                 0.00    0.31     0.46             0.35
## xperchr                            0.31    0.00     0.12             0.77
## xpercell                           0.46    0.12     0.00             0.34
## log10GeneticSize                   0.35    0.77     0.34             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.4.2 totalArmGains



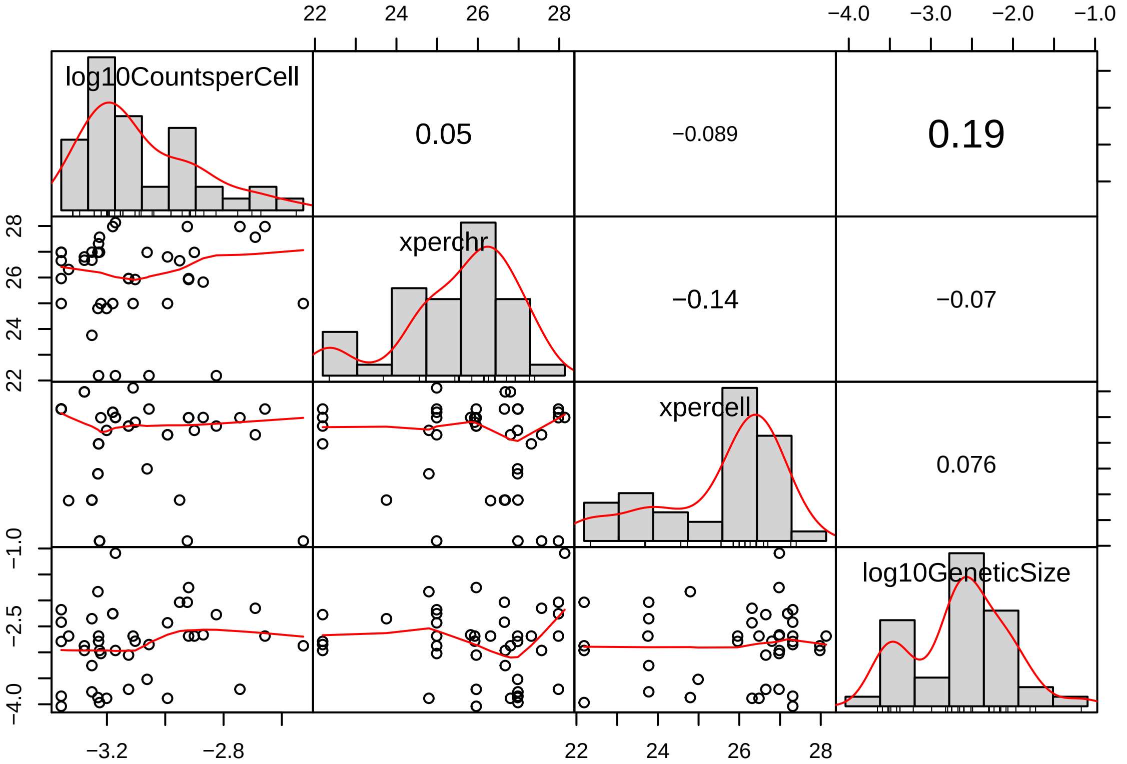
```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                    log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell               1.00    0.03     0.00            -0.03
## xperchr                          0.03    1.00    -0.07            -0.16
## xpercell                         0.00   -0.07     1.00            -0.41
## log10GeneticSize                -0.03   -0.16    -0.41             1.00
## Sample Size
## [1] 28
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                    log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell               0.00    0.86     1.00             0.87
## xperchr                          0.86    0.00     0.73             0.40
## xpercell                         1.00    0.73     0.00             0.03
## log10GeneticSize                 0.87    0.40     0.03             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.4.3 totalWholeLosses



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                    log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell               1.00   -0.02     0.21             0.13
## xperchr                         -0.02    1.00     0.08             0.02
## xpercell                         0.21    0.08     1.00            -0.11
## log10GeneticSize                 0.13    0.02    -0.11             1.00
## Sample Size
## [1] 104
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                    log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell               0.00    0.82     0.03             0.19
## xperchr                          0.82    0.00     0.45             0.83
## xpercell                         0.03    0.45     0.00             0.27
## log10GeneticSize                 0.19    0.83     0.27             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

2.4.4 totalWholeGains



```
## Call:corr.test(x = analyze, use = "pairwise", method = "pearson",
##     adjust = "none", alpha = 0.05)
## Correlation matrix
##                      log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell                 1.00    0.05    -0.09             0.19
## xperchr                            0.05    1.00    -0.14            -0.07
## xpercell                          -0.09   -0.14     1.00             0.08
## log10GeneticSize                   0.19   -0.07     0.08             1.00
## Sample Size
## [1] 42
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                      log10CountsperCell xperchr xpercell log10GeneticSize
## log10CountsperCell                 0.00    0.75     0.58             0.23
## xperchr                            0.75    0.00     0.37             0.66
## xpercell                           0.58    0.37     0.00             0.63
## log10GeneticSize                   0.23    0.66     0.63             0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```