Systems Genetics 2021

Final Project

Adi Sivan Michal Alayev

1. Defining the Task

We focused on the collection of traits related to the **proliferation of stem and progenitor cells**.

We selected the next phenotypes:

1. **206: Immune function: Proliferative capacity in vitro of bone marrow stem and progenitor cells from young mice, proliferation of lin-Sca1++ cells from 8-week-old mice in response to KL, flt3L, and TPO (n of cells per 50 input cells after 5 days of culture) [n]**, Henckaerts E, Geiger H, Langer JC, Rebollo P, Van Zant G, Snoeck HW, 2002
2. **207: Immune function: Hematopoietic stem cells and progenitor cells, numbers of lin-Sca1++ cells at 8 weeks [n]**, Henckaerts E, Geiger H, Langer JC, Rebollo P, Van Zant G, Snoeck HW, 2002
3. **210: Immune system and blood, hematopoietic stem cells and progenitor cells (lin-Sca1++ c-kit+ cells, LSK) from femur [n]**, Henckaerts E, Langer JC, Snoeck HW, 2004

We chose using gene expression from the next datasets, as relevant intermediate traits:

1. **Myeloid**, GEO database, Access ID: GSE18067
2. **Blood stem cell**, GEO database, Access ID: GSE18067

These tissues were selected since they are directly related to our phenotypes of interest, the gene expression in these tissues is more likely to be associated with the proliferation of the stem cells and the Myeloid progenitors that differentiate from them.

2. Gene Expression Preprocessing

The preprocessing is done in the script **Q2.py**

The data we use can be found in **GSE18067_series_matrix.txt** (both myeloid and blood stem cells) and was downloaded as normalized data. The preprocessing step outputs two files: **Myeloid_preprocessed.csv** and **Stem_preprocessed.csv**. Log is presented in **Q2_log.txt**.

Steps that were taken for each gene expression dataset:

- Gene expression data was averaged across different individuals of the same strain.
- Removing genes with no identifier.
- Removing genes with no positions in the genome.
- Removing rows with low maximal value – we chose to remove the lower 70^{th} percentile of the data.
- Removing rows with low variance – here too, we chose to remove the lower 70^{th} percentile of the data.
- In case of multiple rows (probes) for the same gene, we calculated their average.

- Filtering by neighboring loci was done in the eQTL and the QTL analysis parts.

3. eQTL Analysis

Myeloid gene expression dataset

In this dataset there are 24 different BXD strains, and 1309 genes (after preprocessing).
We kept the data of only those 24 strains in the genotypes file (from EX2), and afterwards filtered loci that have exactly the same information as neighboring loci across the 24 BXD strains.
The genotypes were converted to numeric values, where B=0, H=1, D=2. This step created a file called **numeric_filtered_genotypes_Myeloid.csv**, which contains 818 loci out of 3796 in the original genotypes file.

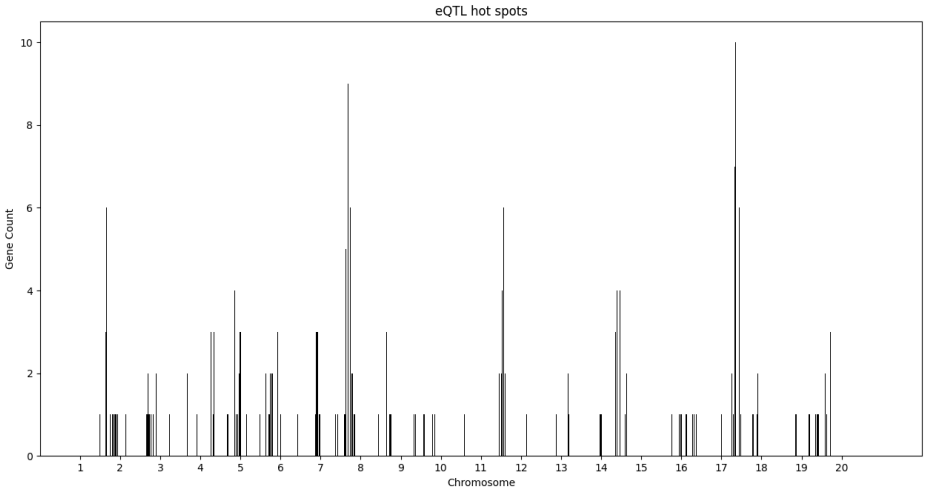
We performed an association test using regression on all genes in the gene expression input file (**Myeloid_preprocessed.csv**), using each of the SNPs in the filtered genotypes file. The amount of tests performed is 1309 genes x 818 SNPs = 1,070,762.
The p-values of the tests were saved in **eQTL_results_Myeloid.csv** (uncorrected) and in **eQTL_results_corrected_Myeloid.csv** (corrected using Bonferroni multiple testing correction).

The SNPs with significant p-value after multiple testing correction were devided to cis-acting trans-acting eQTLs, when an eQTL is considerd to be cis-acting when the SNP and gene are located on the same chromosome and the distance between the SNP and gene edges is 2Mbp or less.

For the significance level 0.05, and due to the number of tests that is used for correction, the p-value before correction should be smaller than 4.67x10^-8 in order to be statistically significant. This is a very low threshold, so it is expected that we will find only a few eQTLs.

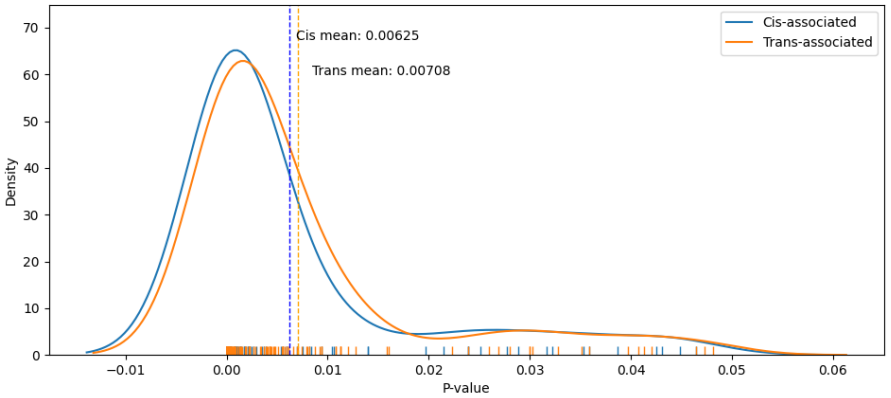
Results:
We found 246 different significant eQTLs (after multiple testing correction), where 109 of them are cis-acting and 137 are trans-acting. All trans-acting eQTLs (except for one) are located on the same chromosome as their associated gene, but the distance between the SNP and gene edges is more than 2Mbp. We found only one trans-acting eQTL that is located on a different chromosome than its associated gene.
130 different genes were associated with at least one SNP, while 136 SNPs were associated with at least one gene, which means that there were SNPs associated with more than one gene.
The description of all eQTLs (gene, SNP, chromosome, and corrected p-value) is present in **all_eQTLs_Myeloid.csv.**

The number of genes associated with each eQTL across the genome is presented in the plot below:



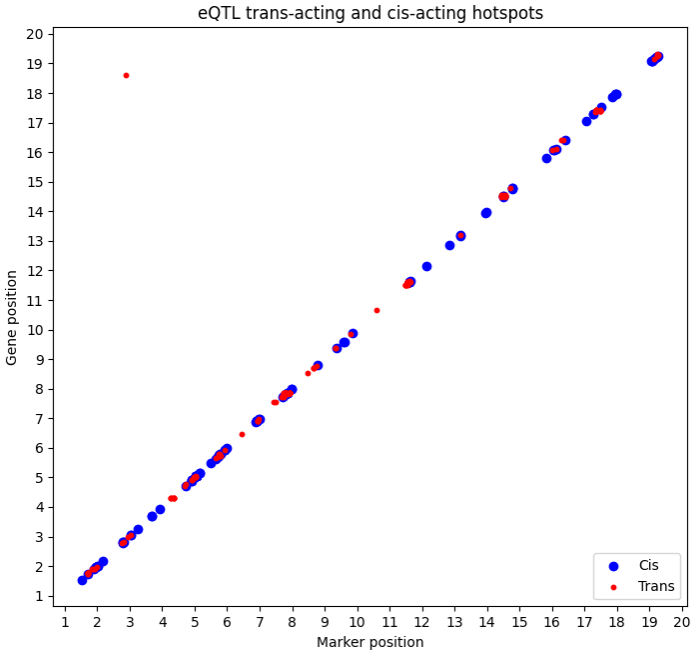
We can see significant hotspots in chromosomes 1, 7, 11, and 17, with SNPs that are associated with 6-10 genes.

The distribution of association P-values (corrected) between cis-associated genes and trans-associated genes:



The distributions are very similar, with a little smaller mean for the cis distribution. This is unexpected, since we assume that cis-acting eQTLs have a stronger association to the associated genes then trans-acting eQTLs. A possible explanation is that most of the trans-acting eQTLs (all but one) are located on the same chromosome as the gene (as shown in the scatter plot below), therefore the association is strong, but the eQTL is considered as trans-acting due to the predefined distance range.

Scatter plot of trans-acting and cis-acting eQTLs across the genome:



It can be assumed that because we used the Bonferroni correction that places a rigid significance threshold when the number of tests is large, trans-acting eQTLs that are at a greater distance from the associated gene have not crossed the threshold and therefore do not appear in the plot.
It is possible that using a less rigid correction would increase the amount of tests with a significant P-value, and thus we would find additional trans-acting eQTLs that won't be located on the diagonal of the plot, but will be located a little further from the gene they affect.

Blood stem cell gene expression dataset

The analysis includes the same steps we performed for the Myeloid dataset.

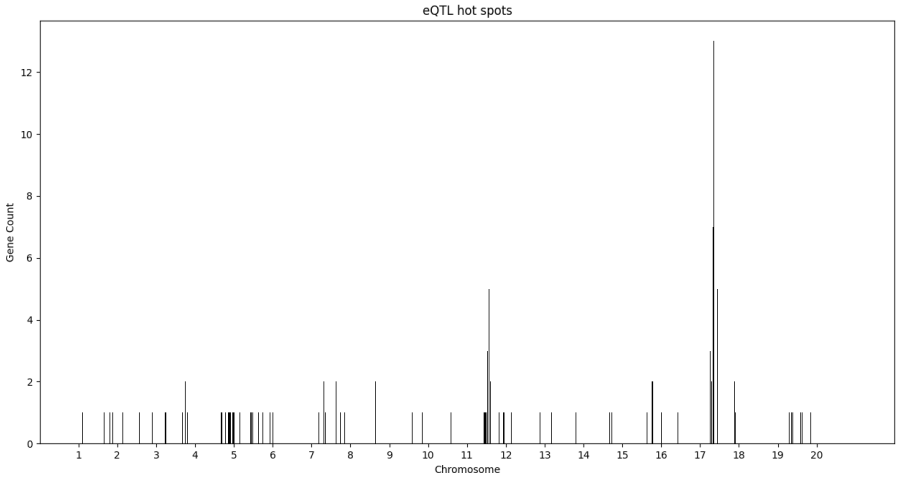
In this dataset there are 23 different BXD strains, and 1309 genes (after preprocessing).
As in the previous gene expression dataset, we kept the data of only those 23 strains in the genotypes file , and afterwards filtered loci that have exactly the same information as neighboring loci across the 23 BXD strains. 808 loci out of 3796 in the genotypes file were left.
The remaining genotypes were converted to numeric values, where B=0, H=1, D=2. This step created a file called **numeric_filtered_genotypes_Stem.csv**.

We performed an association test using regression on all genes in the gene expression input file (**Stem_preprocessed.csv**), using each of the SNPs in the matching filtered genotypes file. The amount of tests performed is 1309 genes x 808 SNPs = 1,057,672 (a little less than for the Myeloid dataset).
The p-values of the tests were saved in **eQTL_results_Stem.csv** (uncorrected) and in **eQTL_results_corrected_Stem.csv** (corrected using Bonferroni multiple testing correction).

For the significance level 0.05, and due to the number of tests that is used for correction, the p-value before correction should be smaller than 4.73x10^-8 in order to be statistically significant. This is almost as the threshold for Myeloid dataset, and is a very low threshold, so it is expected that we will find only a few eQTLs.

Results:
We found 121 different significant eQTLs (after multiple testing correction), where 45 of them are cis-acting and 76 are trans-acting. As for the Myeloid dataset, all trans-acting eQTLs except for one are located on the same chromosome as their associated gene, but the distance between the SNP and gene edges is more than 2Mbp. We found only one trans-acting eQTL that is located on a different chromosome than its associated gene.
84 different genes were associated with at least one SNP, while 77 SNPs were associated with at least one gene.
The description of all eQTLs of this dataset (gene, SNP, chromosome, and corrected p-value) is present in **all_eQTLs_Stem.csv.**

The number of genes associated with each eQTL across the genome is presented in the plot below:

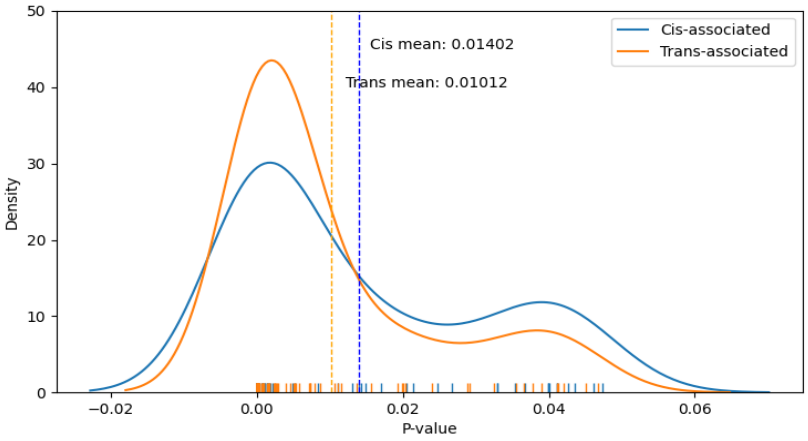


We can see significant hotspots in chromosomes 11 and 17, which were also noticeable in the Myeloid dataset. Further inspection shows that there are 3 SNPs in chromosome 17 that are

associated to a large number of genes, in both Myeloid and blood stem cell datasets, as presented in the table below:

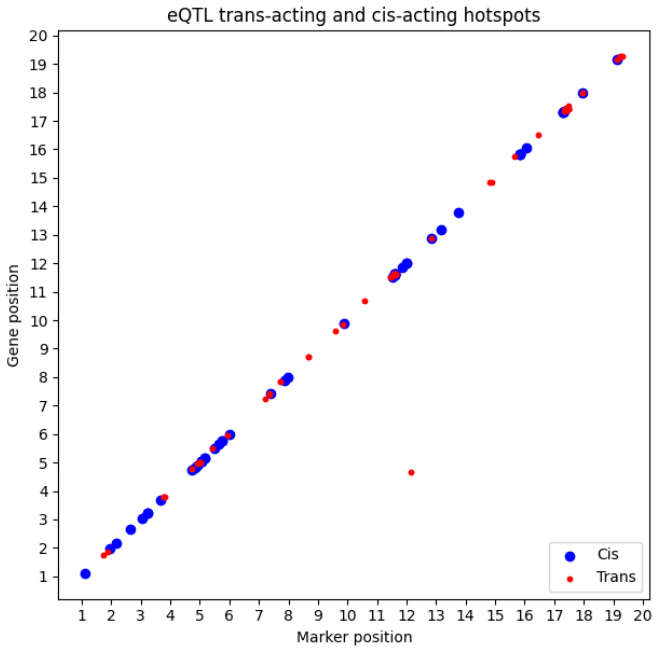
SNP	Number of associated genes Myeloid	Number of associated genes Stem
rs13482947	7	7
rs6242153	10	13
rs13482981	6	5

The distribution of association P-values (corrected) between cis-associated genes and trans-associated genes:



The cis distribution has a smaller mean than the trans distribution, which is unexpected. This can be explained the same way as for the Myeloid dataset distributions: most of the trans-acting eQTLs (all but one) are located on the same chromosome as the gene (as shown in the scatter plot below), therefore the association is strong, but the eQTL is considered as trans-acting due to the predefined distance range (distance of 2Mbp or less between the the SNP and gene edges is considerd to be cis-acting).

Scatter plot of trans-acting and cis-acting eQTLs across the genome:



We see that almost all the eQTLs are located near the gene they are associated with and on the same chromosome, including the trans-acting eQTLs. In addition, there are fewer eQTLs than for the Myeloid dataset.

Log of the code run is presented in **eQTL_and_QTL_analysis_log.txt**.

4. QTL Analysis

We ran a genome wide association test on each of our selected phenotypes, when the input files were **genotypes.csv** and **phenotypes.csv** (we converted the txt files we got in EX2 to csv files).

For each phenotype:

- We first kept only those BXD strains that have a value for the trait, and then kept only these strains in the genotypes file.
- Then we filtered neighboring loci in the reduced genotypes file that have exactly the same information. This step left us a different amount of loci for each phenotype to run the QTL analysis on.
- Finally, we ran a regression association test using each of the SNPs in the filtered genotypes file, defining B=0, H=1, D=2.
- The p-values of the tests were saved in **QTL_results_X.csv** (uncorrected) and in **QTL_results_corrected_X.csv** (corrected using Bonferroni multiple testing correction), when X is the ID of the phenotype as presented in phenotypes.csv.

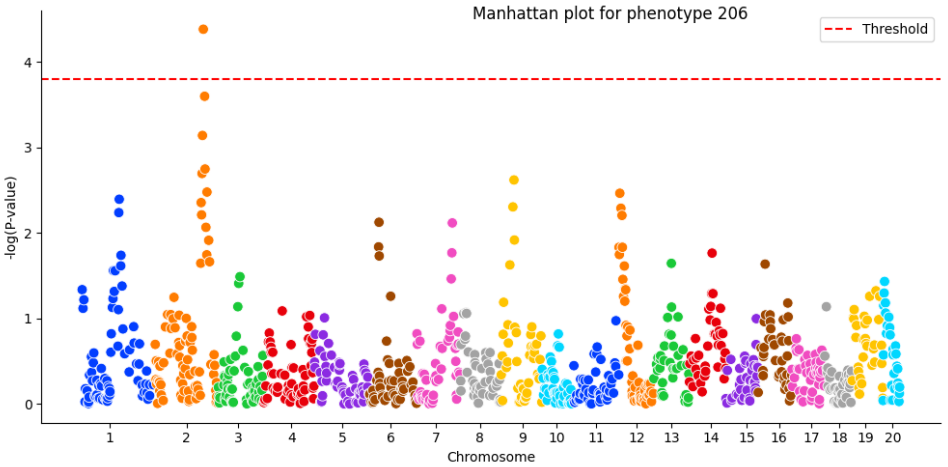
For the significance level 0.05, and due to the multiple testing correction, the p-values before correction should be smaller than 1.32x10^-5 in order to be statistically significant, or above 4.88 in -log10(p-value) scale. This is a very high threshold, so it is expected that we will find only a few QTLs. The threshold after correction is **0.05** for p-value and **1.30** for -log10(p-value).

We found only one SNP that passed this threshold, rs6206791, this QTL was found in chromosome 2 in position 149491612, with corrected p-value of 0.039, and was associated with the phenotype 206. In order to be able to run causality test in the next part of the project, we needed to find more QTLs, so we decided to try different significance levels. We eventually chose 0.15 as the significance level for QTLs, which gave us enough QTLs for performing the causality test.

The results below show the QTLs we found for each of our selected phenotypes, when $\alpha = 0.15$. The threshold of significance was different for each phenotype, since the number of loci (which is equivalent to the number of tests performed) was different for each phenotype after the filtering of the genotypes file, and came into account when performing multiple testing correction.

Results:
For each phenotype we generated a Manhattan plot with the -log10(P-value) of each SNP before correction, specifying the threshold.

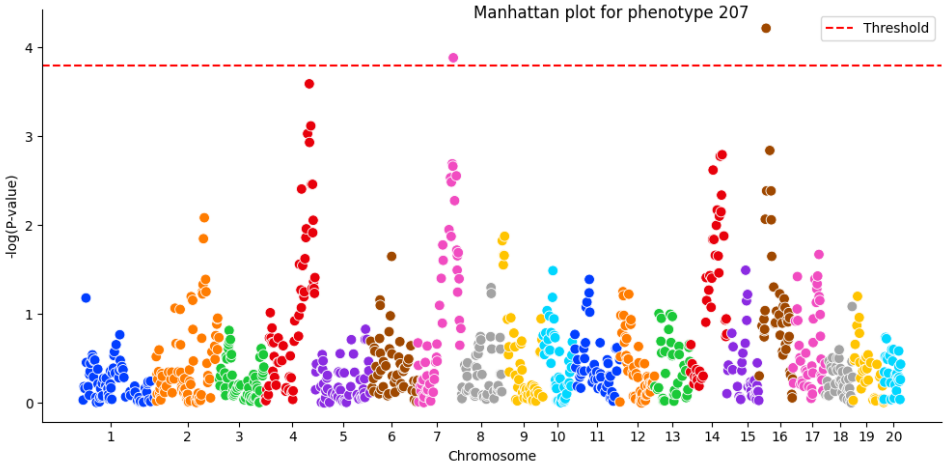
Phenotype 206:
After filtering the loci, 941 loci remained, so the threshold was 0.15/941 = 1.59x10^-4 or 3.80 in -log10(p-value) scale before correction.
We found one significant QTL, rs6206791, located in chromosome 2 in position 149491612, with p-value before correction of 4.13x10^-5 or 4.38 in -log10(p-value) scale.
The corrected p-value was 0.0389, and -log10(p-value) of 1.41.



Phenotype 207:

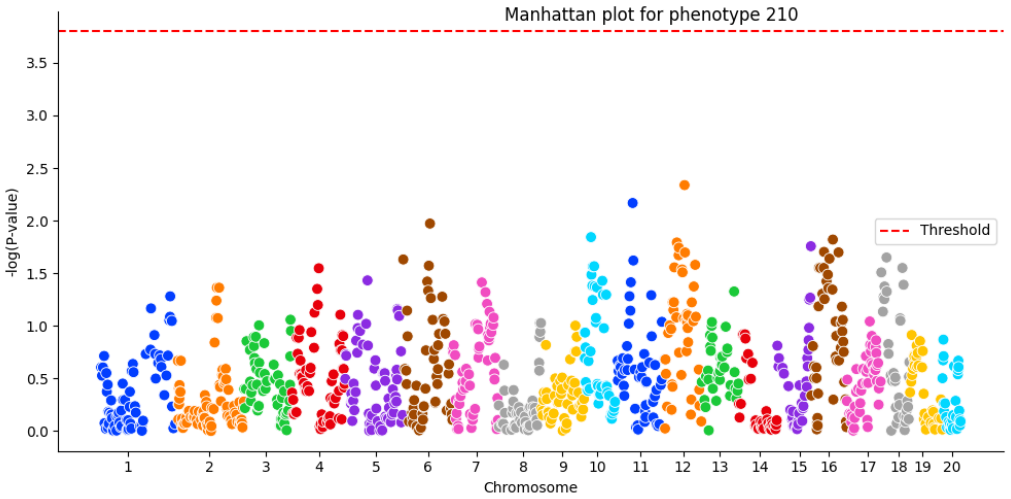
After filtering the loci, 941 loci remained, so the threshold was 0.15/941 = 1.59x10^-4 or 3.80 in -log10(p-value) scale before correction, as for the previous phenotype.
We found two significant QTLs:

1. rs13479470, located in chromosome 7 in position 121842519, with p-value before correction of 1.30x10^-4 or 3.89 in -log10(p-value) scale.
   The corrected p-value was 0.122, and -log10(p-value) of 0.91.
2. rs4165503, located in chromosome 16 in position 27062966, with p-value before correction of 6.06x10^-5 or 4.22 in -log10(p-value) scale.
   The corrected p-value was 0.057, and -log10(p-value) of 1.24.



Phenotype 210:

After filtering the loci, 944 loci remained, so the threshold was 0.15/944 = 1.589x10^-4 or 3.799 in -log10(p-value) scale before correction, very similar to the previous phenotypes.
For this phenotype no significant QTLs were found. We can see that the -log10(p-value) values are a lot lower than the significance threshold. This can be seen as an example of a reduction in the power of the test caused by a Bonferroni correction, since we are performing a large number of tests, or maybe there is really no association between the phenotype and the SNPs that were left after filtration.



Over all, we found 3 QTLs in total for the selected phenotypes.

5. Combine Results

Comparing the QTLs of our phenotypes with the collection of eQTLs (which is done in **Q5.py**), we found out that the SNP rs6206791is eQTL for the Myeloid gene expression, and QTL for the phenotype 206.
Besides this SNP, the other two QTLs we found were not identified as eQTLs.

Since we found eQTLs that were also identified as QTLs for our phenotypes later on, we can conclude that SNPs that are eQTLs can help identify QTLs. Moreover, we found some eQTLs in chromosome 2, and we also found a QTL that is located on that chromosome. So eQTLs can help us focus on specific locations in the genome (such as specific chromosomes) in order to check whether they contain SNPs that are QTLs, instead of performing analysis on the entire genome. This can help prevent loss of power of the tests due to multiple testing correction, since we perform less tests.

In case we limit GWAS to the SNPs that are associated with at least one gene expression trait, meaning limiting GWAS to SNPs that are eQTLs, we would miss out potential QTLs, since in our study we found two more QTLs that were not eQTLs. Hence, from our experience, GWAS should not be limited to SNPs that were identified as eQTLs beforehand, but these SNPs can narrow down the locations for testing.

The file **Q5_results.txt** describes in detail the SNPs that were identified as QTLs and also eQTLs, for each gene expression dataset and each phenotype.

6. Causality analysis

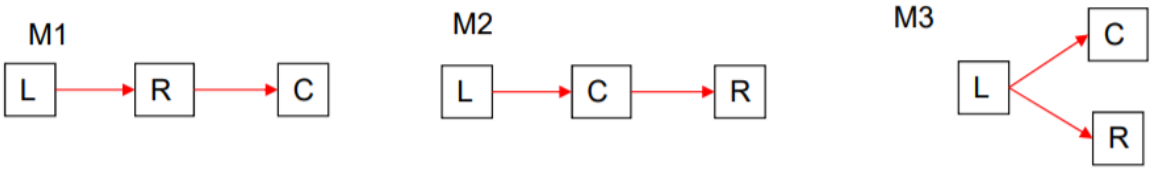
1. Computing the Pairs

From the eQTL and QTL results we chose the ones that passed a p-value threshold, and searched for pairs nearby (cis) loci. The valid pairs were saved to the file **pairs_Myeloid.csv** and **pairs_Stem.csv**

2. Choosing the Best Model

For each pair, we extracted data of the relevant BXD strains, i.e strains which had data both for the gene expression and for the complex trait of the pair.

For each pair, we computed the likelihood of each model based on the extracted data, as described in Lecture 8 pdf (Lecture10b.pdf) slides 21-26. As we denote locus as L, gene as R and complex trait as C the models and the likelihood calculations are described as follows:



$$P(L,R,C \mid \theta_{M1}) = \prod_{i=1}^{n} P(L_i) \cdot P(R_i \mid L_i) \cdot P(C_i \mid R_i)$$

$$P(L,R,C \mid \theta_{M2}) = \prod_{i=1}^{n} P(L_i) \cdot P(C_i \mid L_i) \cdot P(R_i \mid C_i)$$

$$P(L,R,C \mid \theta_{M3}) = \prod_{i=1}^{n} P(L_i) \cdot P(R_i \mid L_i) \cdot P(C_i \mid L_i)$$

We tried to define the best model, based on the AIC criteria discussed in the class. More specifically, we used the Corrected AIC, as described in the next equations:

$$\mathrm{AIC} \;=\; 2k - 2\ln(\hat{L})$$

$$\mathrm{AICc} \;=\; \mathrm{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

These equations are taken from Wikipedia.

3. Computing Model Significance

In order to compute the significance of the accepted model, we conducted permutation tests as follows (for each pair separately):

Permutation test for the a model of a pair:

1. Repeat 1000 times:
    a. Given a matrix with the relevant BXDs data (3 columns which represent L,R,C and the row for each BXD), shuffle the values within the R in order to break the connection between the gene expression and the locus and complex trait factors; then shuffle the C column, in order to break the connection left between the trait and the locus.
    b. Compute the most likely model (the same way it was done on the original data).
2. Return `(1000 - n)/1000` when `n` is the number of times which the original model was chosen.

4. Final Results

As described above, we used p-value thresholds in order to filter the eQTLs and QTLs which were used later in the pairs. We decided to pick the same p-value for both QTL and eQTL in order to be consistent.

We tried three different p-values: **0.15**, **0.1** and **0.05**. The reason for that was that we had only one significant QTL for p-value of 0.05, and we wanted to try our model estimation for a larger number of pairs.

Mind that all the pairs that were accepted include eQTLs from the **Myeloid** tissue, since the Blood Stem Cells didn't produce significant enough p-values.

For p-value of 0.15 there were found 7 pairs with the next models:

```
1.
For Myeloid , for the locus rs6206791 , the gene Cdk5rap1 and the
phenotype 206 the most likely model is Model 3
     Model pVal is 0.000

2.
For Myeloid , for the locus rs4165676 , the gene Slc15a2 and the
phenotype 207 the most likely model is Model 3
     Model pVal is 0.000
     Mind that rs4165676 is the eQTL locus. You might also want
to consider the nearby QTL locus: rs4165503

3.
For Myeloid , for the locus rs6206791 , the gene Kif3b and the
phenotype 206 the most likely model is Model 3
     Model pVal is 0.584

4.
For Myeloid , for the locus rs3674465 , the gene A530023O14Rik
and the phenotype 207 the most likely model is Model 3
     Model pVal is 0.648
     Mind that rs3674465 is the eQTL locus. You might also want
to consider the nearby QTL locus: rs13479470

5.
For Myeloid , for the locus rs3674465 , the gene Trim30 and the
phenotype 207 the most likely model is Model 3
     Model pVal is 0.000
     Mind that rs3674465 is the eQTL locus. You might also want
to consider the nearby QTL locus: rs13479470

6.
For Myeloid , for the locus rs6280792 , the gene A530023O14Rik
and the phenotype 207 the most likely model is Model 3
     Model pVal is 0.631
     Mind that rs6280792 is the eQTL locus. You might also want
to consider the nearby QTL locus: rs13479470
```

7.
For **Myeloid ,** for the locus **rs6280792 ,** the gene **Trim30** and the phenotype **207** the most likely model is **Model 3**
Model pVal is **0.000**
Mind that rs6280792 is the eQTL locus. You might also want to consider the nearby QTL locus: **rs13479470**

For p-value of 0.1 only the above **three first results** were accepted (with the same model p-values)

For p-value of 0.05 only the above **first result** was accepted (with the same model p-value).