Final Project

Maggi Plutov 209223155

Shani Daniel 315231902

1. **Defining the task:**
   We chose to focus on cocaine response traits. We specifically chose the following phenotypes - "*Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]"* and "*Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]"*. We wanted to look at the same time range, sex and activity type, when the difference is the number of injections (1 vs 2), to look for interesting observations.
   We used gene expression from both blood stem cells and liver tissue as relevant intermediate traits (GEO accession IDs: GSE17522, GSE18067).
   The intermediate traits were chosen from the following considerations:

   - **Liver:** Cocaine is primarily metabolized in the liver, thus the liver tissue should be affected by cocaine exposure.
   - **Blood stem cells:** Research has shown that cocaine exposure can lead to changes in hematopoiesis (the process of blood cell formation) and may affect the differentiation and function of blood stem cells. For example, some studies have indicated that cocaine exposure could potentially disrupt normal bone marrow function, leading to alterations in the production of blood cells.
     Thus, we should expect blood stem cells to be affected as well by cocaine exposure.

2. **Gene expression data preprocessing:**

   - We downloaded the normalized data from GEO using the Access ID's provided in the task, for liver and blood stem cells intermediate traits.
   - We combined the two datasets and merged the data and annotation (metadata) information to create our input matrix as stated in the task. The rows of the matrix are gene symbols, and the columns are BXD strain names.
   - We removed rows with no gene identifier (no gene symbol in our case).
   - We removed rows with low maximal value, keeping the top 50% with the highest maximal value (due to the very large number of genes).
   - Similarly, we kept the 50% with highest variance from all the data (before removing low maximal values), and then only kept the rows that are within these both 50% highest maximal value and variance.
     In the liver dataset, we end up with a threshold of 0.3645 as maximal value and 0.0287 threshold for the variance of each row.
     In the blood stem cells dataset, we end up with a threshold of 6.9078 as maximal value and 0.00491 threshold for the variance.
     Thus only rows that meet both these criteria are kept in the final matrix.

- We averaged the data of rows (probes) with the same gene identifier (gene symbol) and thus have 1 row for each gene.
- We filtered neighboring loci in the genotypes file after filtering the BXDs that are given information on in both the current dataset and the genotypes.

After performing all these reprocessing steps we were left with 5,103 genes & 41 different BXD strains in the liver matrix, and 15,993 genes & 25 different BXD strains in the blood stem cells matrix. It corresponds with the information we can find in GEO that states there are 41 and 25 different BXD strains in the liver and blood stem cells datasets respectively.

3. **eQTL analysis:**

We performed a regression test on all genes in the liver final matrix and blood stem cells final matrix separately, using each of the SNPs in the preprocessed genotype file. To define significant p-values we used False Discovery Rate (FDR) correction as we used it in assignment 3, that is uniformly more powerful than the Bonferroni correction and should provide less false-negative eQTLs. The regression combined with FDR correction of the p-values, allowed us to achieve our final collection of eQTLs (eQTLs with corrected P-value < 0.05).

After excluding genes with weak association for all SNPs, we were left with 791 relevant genes in the liver dataset and 900 relevant genes in the blood stem cells dataset.

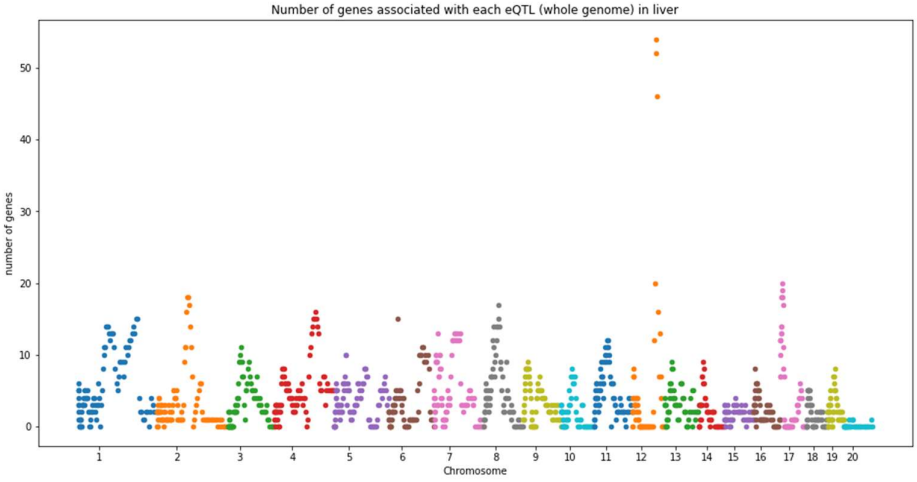
We ended up with 4817 different eQTLs with the liver dataset and 2633 different eQTLs with the blood stem cells dataset. We can observe that despite the much more genes in the processed blood stem cells dataset (around x3) than the processed liver dataset, the number of genes in the liver that seem to be associated with our collection of SNPs is close to the number of genes in the blood stem cells (considering the BXD strains we have information on and the data we obtain).

Also, there are approximately double significant eQTLs found with the liver dataset than the blood stem cells dataset. We can deduct that the SNPs we have probably have more connection to genes in the liver than the blood stem cells.

We can also observe the distribution of number of genes associated with a given eQTL.

In the following plots (in the next page), the x-axis represents the location of each eQTL (SNP) across the entire genome and the y-axis represents the number of genes it's associated to.

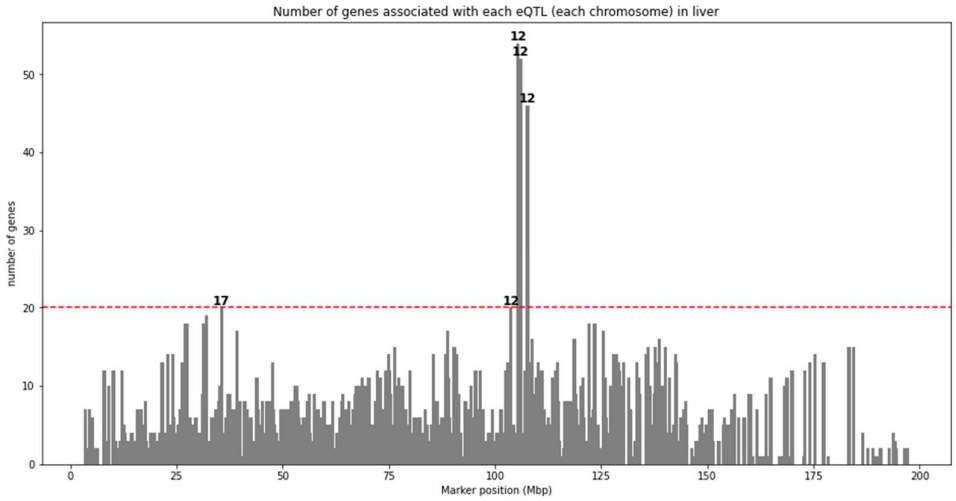
In the liver dataset – we can observe an accumulation of genes in chromosomes 1,2,5,6,7,12,13 and 17, with the majority located on chromosome 12 (maximum of over 50 genes). We can spot a very high peak at chromosome 12.



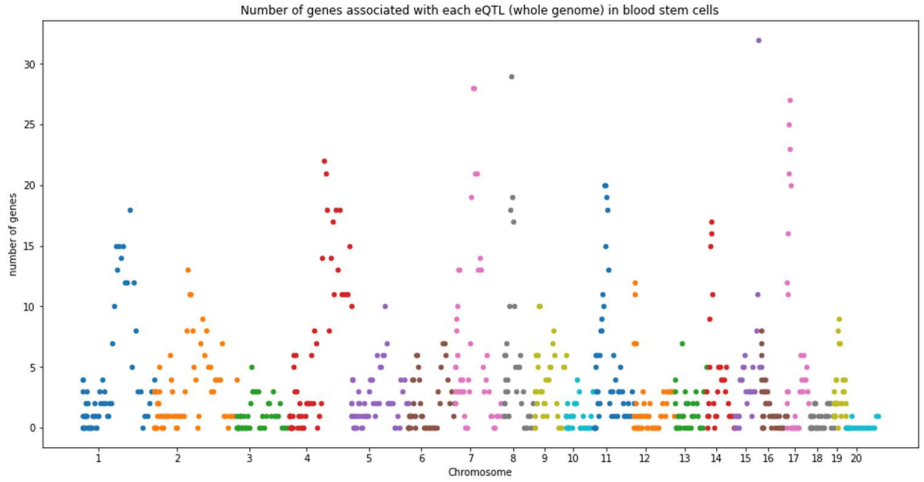
chromosome	Number of associated genes
1	525
2	290
3	211
4	431
5	318
6	236
7	400
8	331
9	248
10	107
11	346
12	321
13	178
14	96
15	99
16	151
17	298
18	89
19	137
20	5

To highlight potential hotspots for further discussion, we generated another plot displaying the number of genes associated with each eQTL, organized by their locations within each chromosome .
The x-axis represents the eQTL location within its chromosome, while the y-axis shows the number of genes. eQTLs with significant number of genes (greater than 20) are labeled with their corresponding chromosome number on top of them.

When we examine the distribution of genes associated with each eQTL within each of the chromosomes, we can identify hotspots, with notable concentrations on chromosomes 17 and 12. Specifically, in the case of chromosome 12, these hotspots are predominantly clustered in the region spanning from approximately 100 to 125 Mbp along the chromosome. That means, chromosome 12 has significantly many genes associated with the SNPs at this area in the genome.

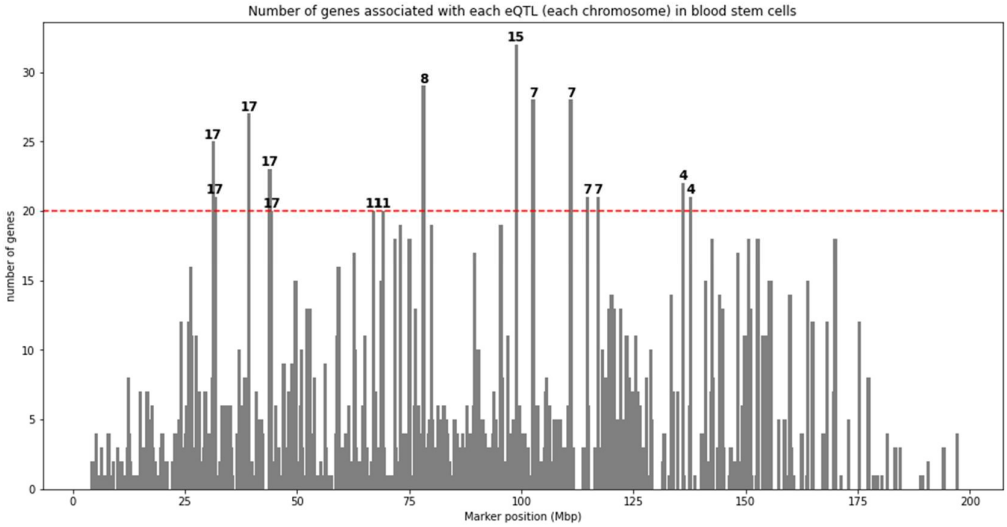


In the blood stem cells dataset – similarly to the liver dataset, we generated two graphs for the number of genes associated with each eQTL across the entire genome and within each chromosome. We observe that in most chromosomes the number of significant genes per SNP reaches a maximum of around 15-20 genes. However, chromosomes 7, 8, 15 and 17 show some significant values, with chromosome 15 having a maximum of over 30 genes, and chromosome 17 relatively many SNPs with over 20 significant genes.



chromosome	Number of associated genes
1	227
2	195
3	44
4	310
5	127
6	85
7	295
8	176
9	132
10	31
11	237
12	89
13	49
14	127
15	120
16	69
17	204
18	25
19	88
20	3

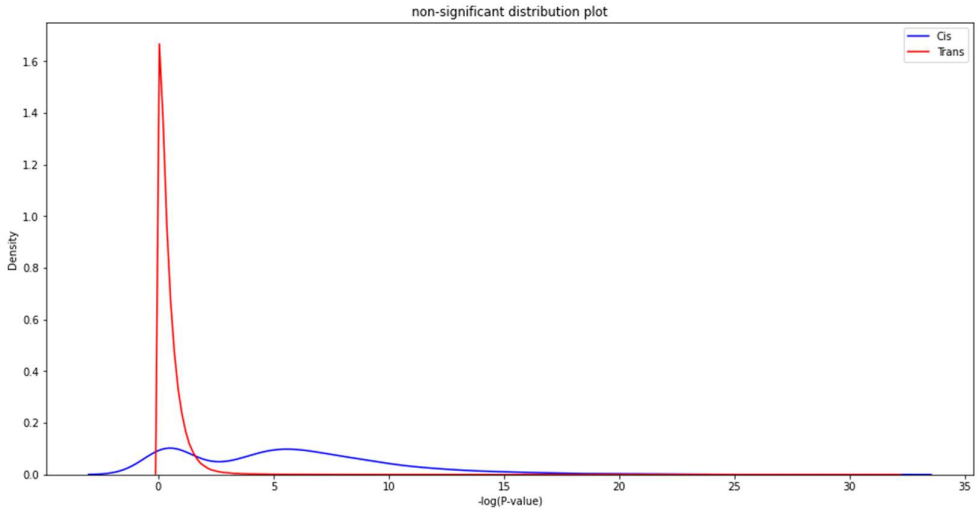
In the second graph, we indeed see potential hotspots in chromosome 17, but this time they accumulate around the beginning of the chromosome. For chromosome 17, this accumulation occurs between 25 and 50 Mbp. We also see the peaks in chromosomes 7, 8 and 15 as expected and seen in the first graph. In addition, now we can see a small peak in chromosomes 4 and 11 as well, thus there are also many genes (over 20) that are associated with certain SNPs in these chromosomes.



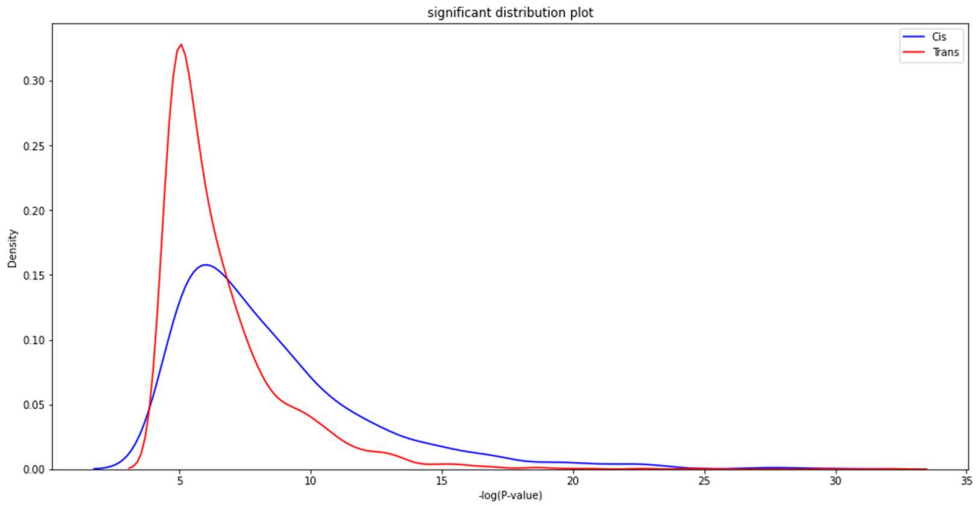
As discussed in class, we would expect that some genomic regions are associated with variation in the expression levels of many transcripts (forming eQTL hotspots), and the expression levels of many transcripts are highly correlated. In our case, we find that most of the associated genes are on chromosome 12 in the liver dataset and chromosomes 7, 8, 15 and 17 in the blood stem cell dataset. In the liver dataset, the number of significant genes in chromosome 12 is surprising with a significant amount of over 50 genes per SNP for some SNPs in that chromosome. In the blood stem cells dataset, there isn't 1 prominent chromosome but some, that also show relatively many genes (over 25).

To further analyze the significant eQTLs, we aspired to also analyzed the cis and trans eQTLs. Due to lack of information on the locations of the blood stem cells genes, we were only able to perform this analysis on the liver eQTLs.
Out of the 4,817 significant liver eQTLs, we found that 1,200 eQTLs are cis-acting, and 3,563 eQTLs are trans-acting. The rest 54 eQTLs had no gene location information.

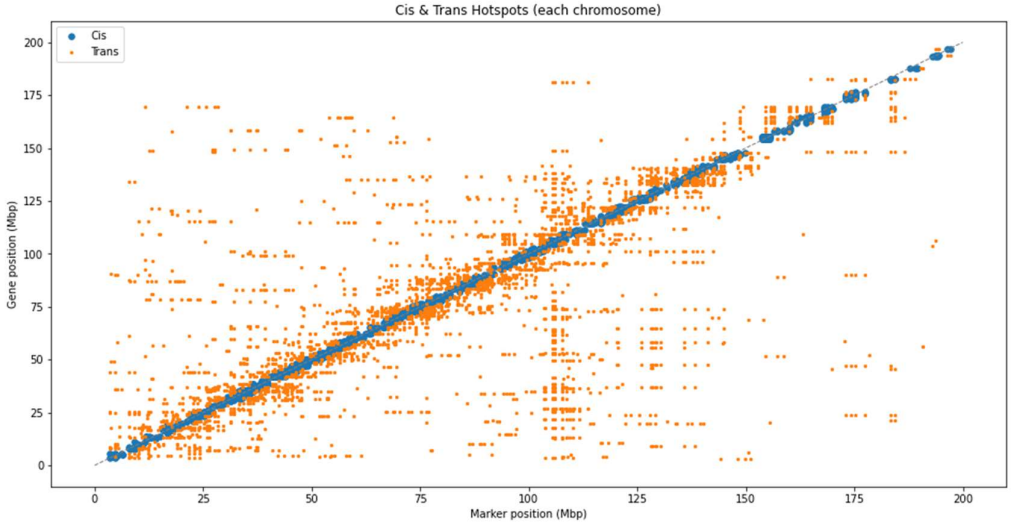
First, we created a cis and trans distribution of the eQTLs. The y-axis shows the density of eQTLs in the liver tissue as a function of the -log(p-value). We can observe that, when looking at all the eQTLs (significant & non significant), that the he -log(P-values) of cis-acting eQTLs are generally higher than trans-acting eQTLs, meaning they are mostly more significant. We can also clearly see that most trans-acting eQTLs are not very significant (their -log(P-value) is very close to zero).



Second, we created the same density plot only for the liver's significant eQTLs, as presented below (in the next page). We can observe an interesting result -
When looking only at the significant eQTLs, we can also see the -log(p-values) of cis-acting significant eQTLs is still slightly higher than trans-acting ones. this aligns with what was discussed in class, given that cis eQTLs tend to have larger effects than trans eQTLs.

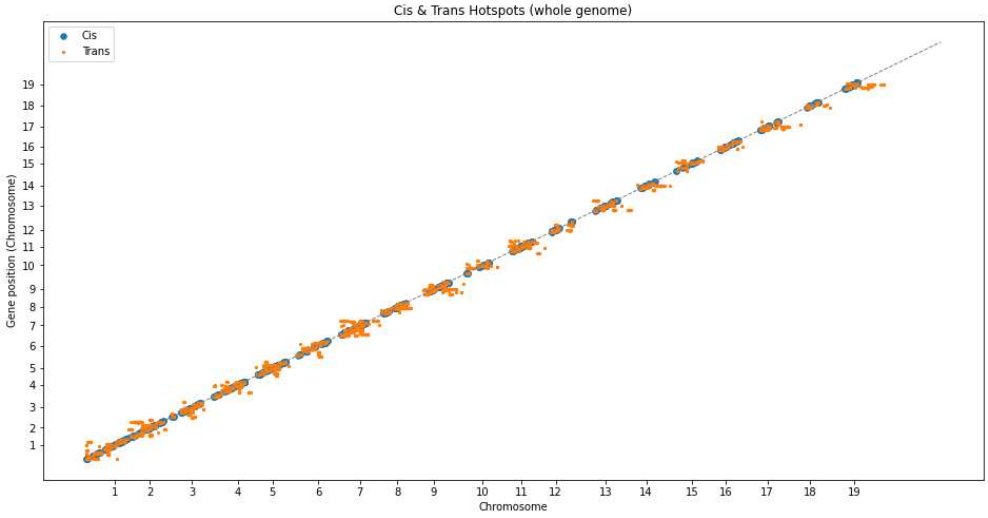


Lastly, we created a cis & trans hotspots plot as seen in lecture -



As expected, the cis-acting eQTLs are on the diagonal, and we got a few vertical lines. We can see an accumulation of vertical lines around 100-125Mbs which may indicate a hotspot. This accumulation corresponds with the significant peak in the former liver graph we presented – number of genes associated with each eQTL, where the peak is also at 100-125Mbp, at chromosome 12. It demonstrates the large number of genes significant with SNPs in that area.

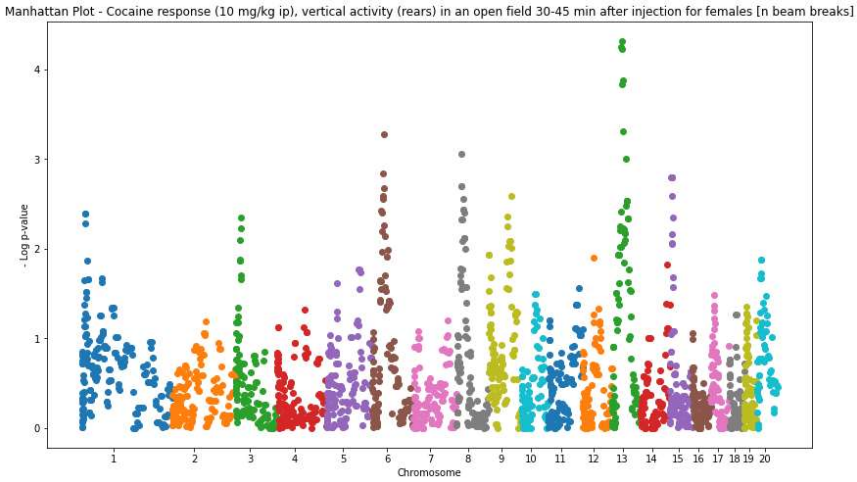
In addition, to deepen the analysis of this locations plot, we created a second plot with the locations by chromosome numbers -

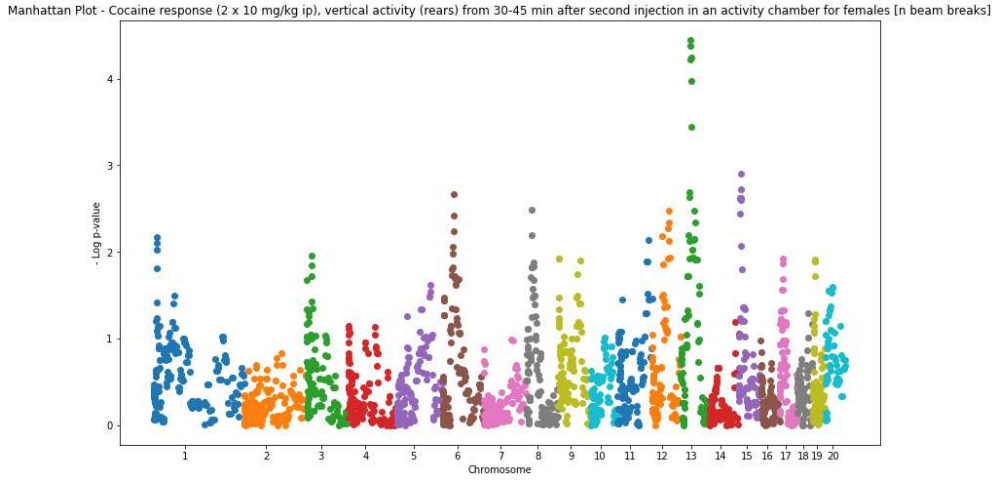


It is evident from the plot that the cis-acting significant eQTLs align along the main diagonal of the graph as well, as expected. Moreover, the trans-acting eQTLs are mostly situated close to the diagonal. This might be attributed to the definition of the cis-acting threshold or due to the scale correction, but still makes sense considering that associated eQTLs tend to be closer together throughout the genome, and specifically on the same chromosome. Furthermore, as seen in the number of associated genes per chromosome table (of the liver), there are many SNPs with significantly associated genes located in the first chromosomes (chromosomes 1-12). Here we can also see this accumulation of eQTLs in the beginning of the graph (blue and orange dots), and less dots (significant eQTLs) after chromosome 13.

4. QTL analysis:

In this part, we ran GWAS test on each of our selected phenotypes. As we stated, we chose phenotypes that are related to cocaine response. We have 2 different phenotypes.
After running linear regression (similarly to assignment 2) and obtaining the p-values, we also used FDR correction to determine the significant QTLs. In this part, because of the small number of phenotypes (2), and the nature of FDR correction, we decided to run the FDR correction on each phenotype separately, considering all the representative SNPs. After this multiple testing correction, we ended up with 8 significant QTLs. That means, there are 8 pairs of SNP-phenotype that are significantly associated to each other, thus probably that SNP has an influence on that phenotype. All these SNPs are located on chromosome 13.
3 QTLs are found in the first phenotype (1 injection), and 5 QTLs in the second phenotype (2 injections). The 3 different significant SNPs (QTLs) of the first phenotype are equal to 3 of the QTLs of the second phenotype. The 3 SNPs that are significant to both phenotypes are – rs3688040, rs3702220, and rs13481905. The 2 SNPs significant only in the second phenotype (2 injections) are rs3722797 and rs6304752.
All the SNPs are also pretty close nearby, all around 80Mbp in the genome.
To further deepen on the results of each phenotype, we created Manhattan plots as in assignment 2:





We can see that the plots are very similar. It makes sense due to the close nature of the 2 phenotypes. The only difference between the phenotypes is the number of injections (1 vs 2). In both plots we see a peak at chromosome 13, that shows a -log p-values of over 4 (p-values < 0.0001). this lines with the significant QTLs we got on chromosome 13. The significant QTLs in chromosome 13 which are significantly associated with both phenotypes are visible, also in the Manhattan plots. We can see that the rest of the chromosomes don't show such high peaks (in relation to chromosome 13), and that is probably why after the FDR correction there are no significant QTLs recognized within them.

5. Combine results:

In this section, we conducted a comparison between the QTLs of our phenotypes and a collection of eQTLs from liver and blood stem cells datasets. To achieve this, we began with the significant eQTLs identified through our gene expression data analysis, which allowed us to narrow down the pool of SNPs for subsequent QTL analysis.

Since the correction for False Discovery Rate (FDR) is performed separately for each phenotype and is influenced by the number of tests (i.e., the number of SNPs), this narrowing down of data was expected to potentially highlight some previously non-significant results.

Furthermore, as mentioned earlier, we anticipated a connection between the expression patterns of genes in the liver and blood stem cells and the response to cocaine, specifically in terms of vertical activity. Consequently, we hypothesized that there would be overlap in the significant SNPs between these datasets.

To assess whether gene expression data indeed adds value to the identification of QTLs, we conducted QTL analysis on 3 SNPs datasets:

1) The union of significant SNPs from the eQTL analyses of both liver and blood stem cells.
2) The intersection of significant SNPs from the eQTL analyses of both liver and blood stem cells.
3) Separately, within each set of significant SNPs from the eQTL analyses of liver and blood stem cells.

For further discussion, we summarized the results in a table of triplets, each containing different combinations of SNP-gene-trait for each of the sets:

Tissue	SNP	Gene	Trait	Chromosome
Blood stem cells	rs13481905	2700017A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	13
		Glrx1		
	rs13481905	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
	rs3722797	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
		Hspg2		
liver	rs13481905	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs13481905	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs3722797	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs6304752	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		

As previously mentioned, our analysis in section 3 yielded 4,817 significant eQTLs (involving 1,112 SNPs) in the liver dataset, with the majority located on chromosome 12. In the blood stem cells dataset, we identified 2,633 significant eQTLs (involving 608 SNPs), primarily on chromosomes 7, 8, 15, and 17.

Our analysis in section 4 on 2320 SNPs, we identified 8 significant QTLs (involving 5 SNPs), of which the phenotype "Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]" had 3 significant QTLs (SNPs), and the phenotype "Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after the second injection in an activity chamber for females [n beam breaks]" had 5 significant QTLs (SNPs). Most of these QTLs were located on chromosome 13.

Interestingly, two of these SNPs, namely rs3688040 and rs3702220, were not found in the list of significant eQTLs in either the liver or blood stem cells datasets, yet they were significant in both phenotypes (total of 4 QTLs). Consequently, these two SNPs were excluded from this section and not considered in the analysis. We compiled the triplets (involving the same SNP) prior to refining the collection of SNPs in the table below.

As observed, we have identified a total of 5 QTLs involving 3 SNPs, with 2 of these SNPs being common to both datasets. In aggregate, we have a total of 19 triplets. Specifically, we have 5 triplets involving 1 SNP (rs13481905) that are shared between both datasets, and these relate to the first injection response phenotype. Additionally, we have 14 triplets and 3 SNPs (rs13481905 and rs3722797 found in both datasets, and rs6304752 exclusively in the liver dataset) associated with the second injection response phenotype. All SNPs are located at chromosome 13.

The following tables will present the significant SNPs that emerged after refining the SNP selection criteria and focusing the GWAS solely on DNA variants linked to at least one expression trait:

1) Union -
   After merging significant SNPs from both datasets (totaling 1,226 SNPs), we found identical results for the second injection phenotype SNP. However, we didn't detect significant QTLs for the first injection phenotype. This could be due to FDR correction, which is more effective with larger sets of tests, offering greater statistical power. In this case, the SNP subset may have led to reduced statistical power for the 'rs13481905' SNP, explaining the absence of significant QTLs.

Tissue	SNP	Gene	Trait	Chromosome
Blood stem cells	rs13481905	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	13
		Glrx1		
	rs3722797	2700017A04Rik		
		Glrx1		
		Hspg2		
liver	rs13481905	Arsk		
		Cox7c		
		Glrx		
	rs3722797	Arsk		

		Cox7c		
		Glrx		
	rs6304752	Arsk		
		Cox7c		
		Glrx		

2) Intersection -

In the intersection set of 494 SNPs, 'rs6304752' is excluded as it was only significant in the liver dataset, not in blood stem cells. Both datasets introduced a significant QTL for the first injection response phenotype with the 'rs3722797' SNP , surpassing the FDR correction threshold while remained significant for the second injection response phenotype. However, there were no changes observed for the second injection phenotype.

Tissue	SNP	Gene	Trait	Chromosome
Blood stem cells	rs13481905	2700017A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	13
		Glrx1		
	rs3722797	2700017A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	
		Glrx1		
		Hspg2		
	rs13481905	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
	rs3722797	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
		Hspg2		
liver	rs13481905	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs3722797	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs13481905	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs3722797	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		

3) Separately -

In this section, we conducted separate QTL analyses on the significant SNPs from the liver dataset (a total of 1,112 SNPs) and the blood stem cells dataset (a total of 608 SNPs). The results for the second injection phenotype remained consistent in both datasets. However, for the first injection phenotype, the liver dataset did not yield any

significant SNPs, whereas the blood stem cells dataset introduced a QTL involving the 'rs3722797' SNP.

Tissue	SNP	Gene	Trait	Chromosome
Blood stem cells	rs13481905	2700017A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	13
		Glrx1		
	rs3722797	2700017A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	
		Glrx1		
		Hspg2		
	rs13481905	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
	rs3722797	2700017A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Glrx1		
		Hspg2		
liver	rs13481905	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	
		Cox7c		
		Glrx		
	rs3722797	Arsk		
		Cox7c		
		Glrx		
	rs6304752	Arsk		
		Cox7c		
		Glrx		

In summary, although it was anticipated that gene expression data would enhance the identification of QTLs, it had a minimal impact on the results for the second injection phenotype, suggesting a strong association between this phenotype and the corresponding SNPs and genes – Arsk, Cox7c, Glrx, 2700017A04Rik, Glrx1 and Hspg2.

Conversely, restricting GWAS to DNA variants linked to at least one expression trait added value for the first injection response phenotype. In certain cases, it led to the discovery of significant QTLs and increased the testing power, while in others, this was not observed due to FDR correction limitations.

Nevertheless, we noted a significant correlation between the first injection response phenotype and the genes Arsk, Cox7c, Glrx, 2700017A04Rik, and Glrx1.

Additionally, as expected, all QTLs were located on chromosome 13, highlighting a robust link between SNPs in this chromosome and the observed phenotypes.

6. Causality analysis:

We applied the causality test on the results from parts 3 and 4. Between all the significant QTLs and eQTLs found in these parts, we looked for triplets where the QTL and eQTL are in nearby genomic position (2Mbp or less from each other). We found 35 QTL-gene-phenotype triplets that answer this limitation, and ran the causality test on each triplet.

As we seen in class, there are 3 potential models tested in the causality test:

M1: L → R → C

M2: L → C → R

M3: L → C, L → R

(L – locus, R – gene, C – trait (phenotype)).

We analyzed each triplet by calculating the likelihood for each model (with the formulas seen in class), and choosing the model with the highest likelihood. We also computed the LR value for each triplet (LR = $\frac{L(M_{max})}{\max(\{L(M_i) \mid M_i \neq M_{max}\})}$).

The results are presented in the next page.

The predicted relations (model) for each triplet and it's LR value is presented in the following table:

(*BSC = blood stem cells)

Tissue	SNP	Gene	Trait	Predicted model	LR value
liver	rs3702220	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	M3	90.10919
		Cox7c		M3	137.123
		Glrx		M3	196.7534
		Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	64.21177
		Cox7c		M3	136.0092
		Glrx		M3	172.8373
	rs3722797	Arsk		M3	64.21177
		Cox7c		M3	136.0092
		Glrx		M3	172.8373
	rs13481905	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	M3	309.7296
		Cox7c		M3	471.3289
		Glrx		M3	676.2944
		Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	102.2916
		Cox7c		M3	216.6675
		Glrx		M3	275.336
	rs6304752	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	54.8322
		Cox7c		M3	116.142
		Glrx		M3	147.5906
BSC	rs3702220	2700017 A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	M3	1.127658
		Glrx1		M1	1.040842
		Hspg2		M3	1.162377
		2700017 A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	3.356119
		Glrx1		M3	3.846612
		Hspg2		M3	2.493939
	rs3722797	2700017 A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	3.356119
		Glrx1		M3	3.846612
		Hspg2		M3	2.493939
	rs13481905	2700017 A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	M3	1.625183
		Glrx1		M3	1.384651
		Hspg2		M3	1.67522
		2700017 A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	2.837704
		Glrx1		M3	3.252431
		Hspg2		M3	2.108704
	rs6304752	2700017 A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	M3	2.837704
		Glrx1		M3	3.252431

As stated before, all the significant SNPs are on chromosome 13.

We can see that 34 out of 35 predicted models are M3, meaning that for 34 out of 34 triplets, M3 is the most likely model. For 1 triplet (SNP: 'rs3702220', Gene: 'Glrx1', Trait: 'Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]') , the predicted model is M1 (with a pretty low LR value of 1.040842).
It can be observed from the table that for triplets with genes from the liver tissue, most LR values are relatively high, all above 50 and most of them above 100.
On the other hand, for the blood stem cells genes, the LR values seem to be significantly lower, all of them around 1 to 3.9.
That can mean that the certainty we have for the predicted model being the real model is probably lower in triplets with blood stem cells genes than liver genes.

To further analyze the statistical significance of the predicted models we got, that means our causality hypotheses, we applied a permutation test on 10 chosen triplets.

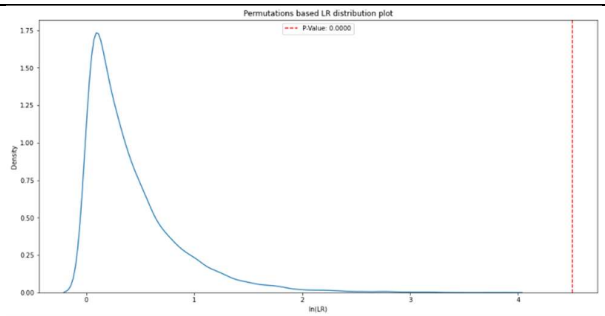
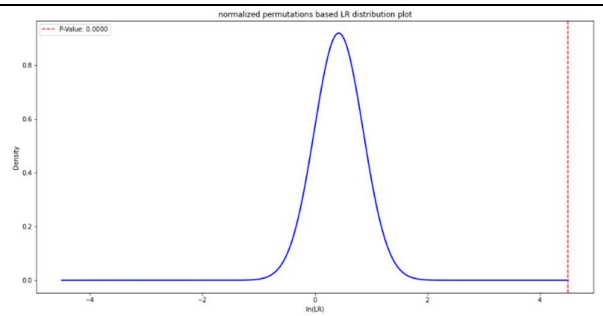
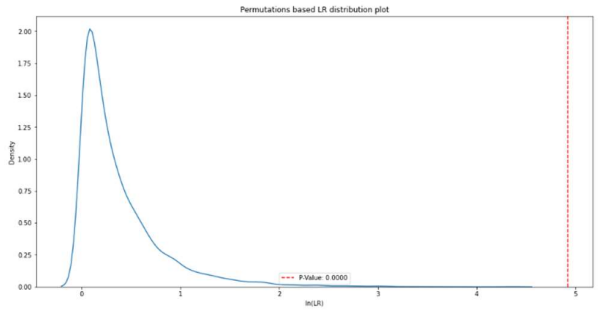
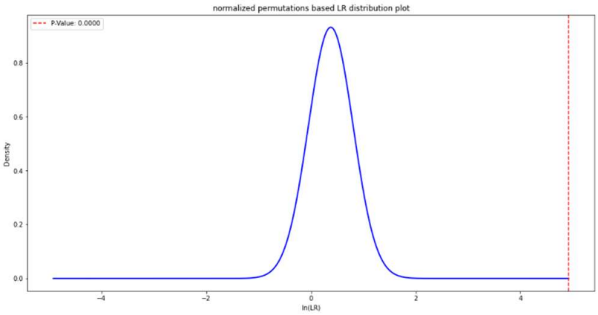
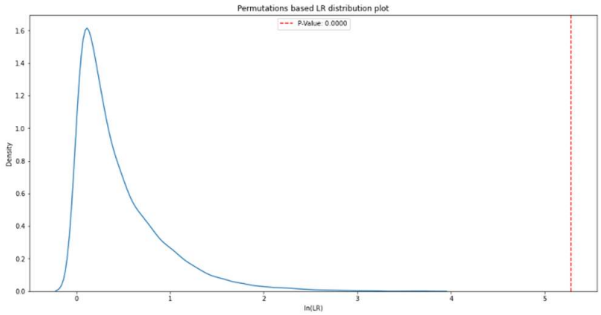
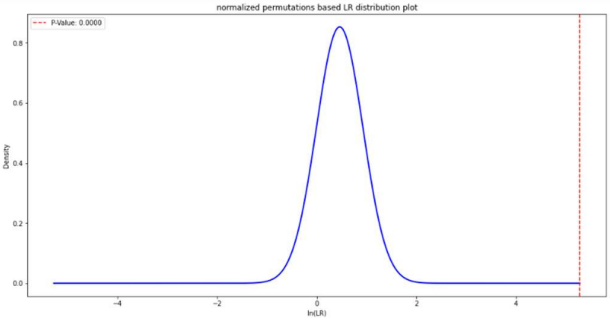
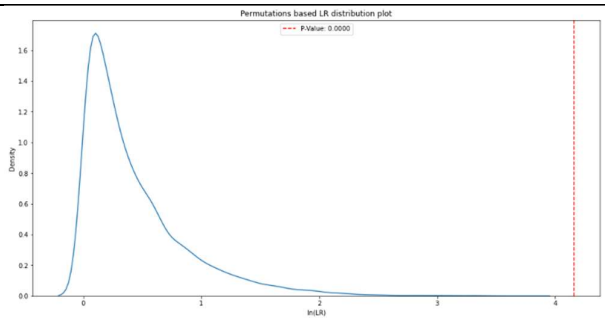
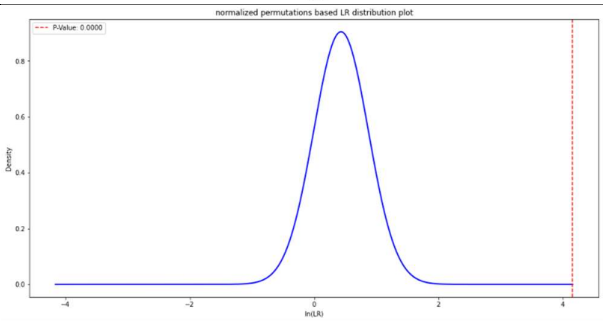
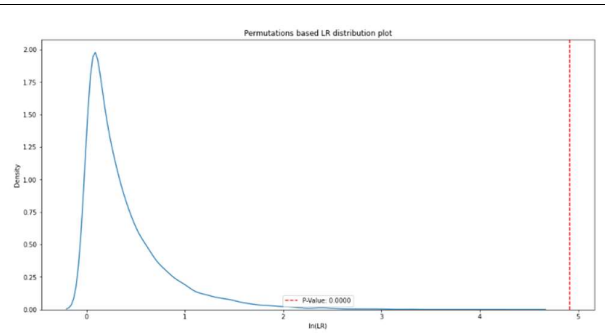
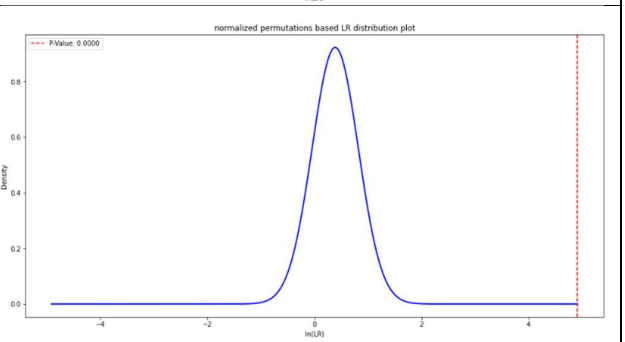
H0: the predicted model is **no closer** to the true model more than the other models
H1: the predicted model is **closer** to the true model more than the other models

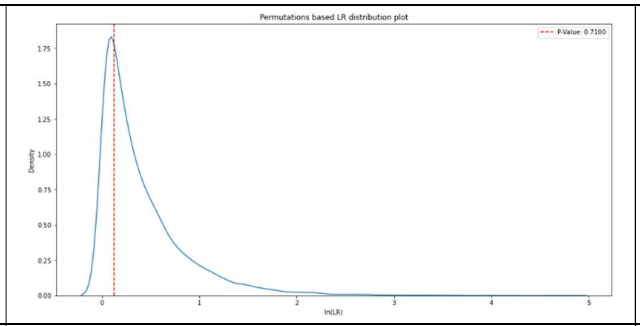
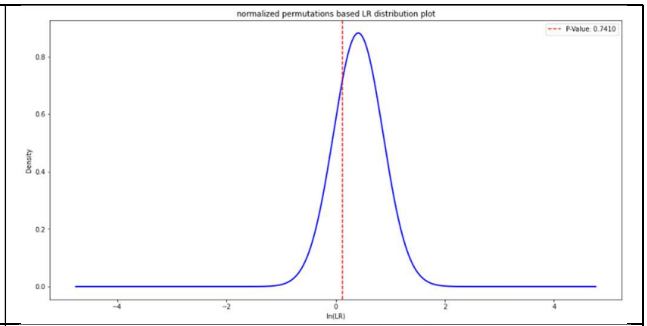
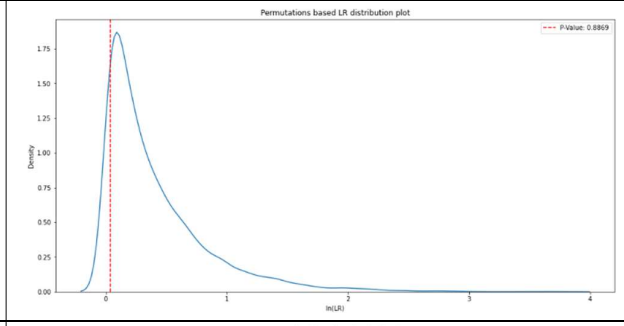
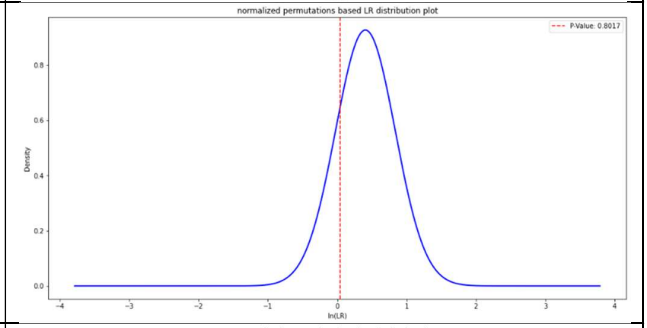
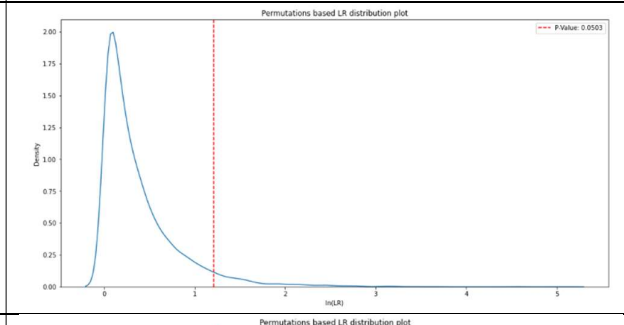
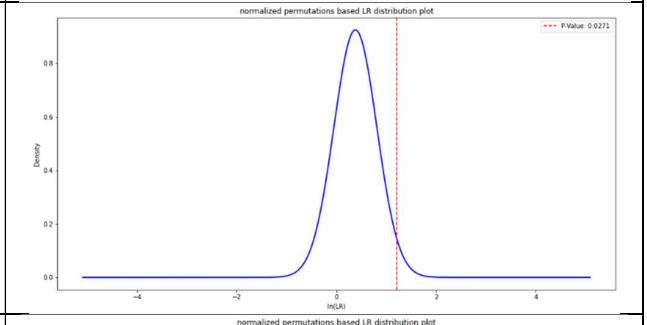
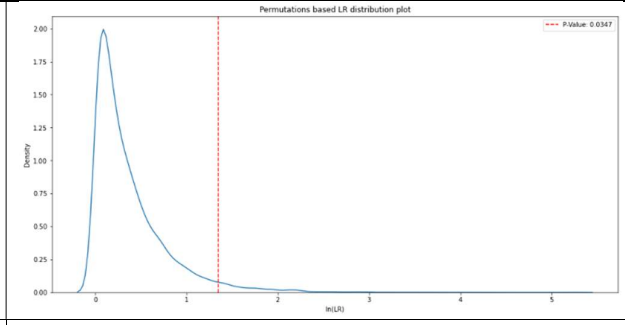
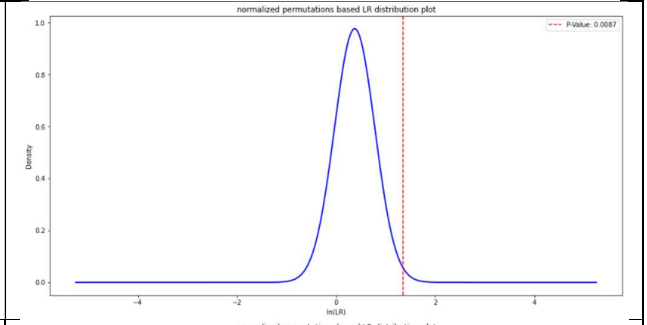
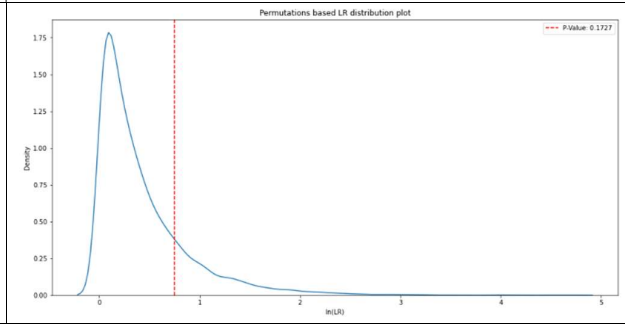
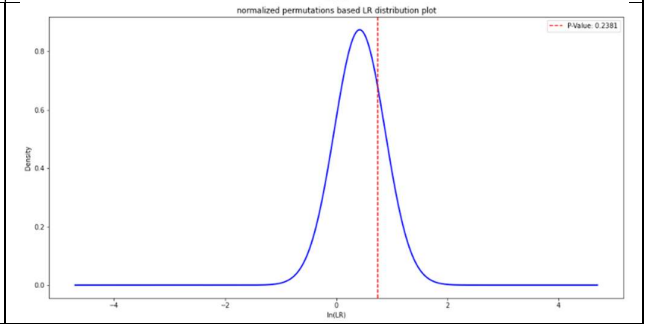
We executed the permutation test as follows:

- For every triplet, we created a table where the columns are L, R, C, as shown in the example in class. We ran the causality model prediction using this table. For the permutation test, we shuffled columns R and C (the gene and trait columns) each separately, to break the former connection between the SNP, gene and phenotype.
- We ran the causality test again on the shuffled triplet and calculated the new LR value.
- We repeated the C&R shuffle, causality test and LR calculation 10,000 times for each triplet.
- We created a graph of the density of ln(LR) values and marked the ln(LR) of the original LR of our predicted model in the graph.
- We calculated the p-value of the causality hypotheses (our predicted model) by checking the number of ln(LR) values we got that are equal to or greater than ln(LR) of our original LR value.
- We also normalized the ln(LRs) graph by calculating their mean and variance for creating normal distribution, and calculated the p-value of the ln(LR) of the original LR from the normal distribution we got.

Here are the results of the permutation test on 10 chosen triplet:

Tissue	SNP	Gene	Trait	ln(LR)	Predicted model	p-value	normalized p-value
Liver	rs3702220	Arsk	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	4.501022	M3	0.0001	0
	rs3702220	Cox7c		4.920879	M3	0	0
	rs3702220	Glrx		5.281951	M3	0.0001	0
	rs3702220	Arsk	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	4.162187	M3	0	0
	rs3722797	Cox7c		4.912723	M3	0	0
BSC	rs3702220	2700017 A04Rik	Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]	0.120143	M3	0.7091	0.74022
	rs3702220	Glrx1		0.04003	M1	0.8945	0.798473
	rs3702220	2700017 A04Rik	Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]	1.210785	M3	0.0509	0.025082
	rs3722797	Glrx1		1.347193	M3	0.0345	0.008346
	rs13481905	Hspg2		0.746073	M3	0.1721	0.227567

Triplet (Tissue, SNP, Gene, Trait)	ln(LR) distribution graph	Normalized ln(LR) graph
Liver, rs3702220, Arsk, Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]		
Liver, rs3702220, Cox7c, Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]		
Liver, rs3702220, Glrx, Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]		
Liver, rs3702220, Arsk, Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]		
Liver, rs3722797, Cox7c, Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]		

BSC, rs3702220, 2700017A04Rik, Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]		
BSC, rs3702220, Glrx1, Cocaine response (10 mg/kg ip), vertical activity (rears) in an open field 30-45 min after injection for females [n beam breaks]		
BSC, rs3702220, 2700017A04Rik, Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]		
BSC, rs3722797, Glrx1, Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]		
BSC, rs13481905, Hspg2, Cocaine response (2 x 10 mg/kg ip), vertical activity (rears) from 30-45 min after second injection in an activity chamber for females [n beam breaks]		

* The vertical red line in the graphs represents the ln(LR) value of the original LR of the predicted model

Due to multiple testing correction (10 tested triplets), we'll define a significant p-value as a p-value lower than 0.005 (0.05/10, Bonferroni correction).
We can see that as we deducted before, the causality hypotheses at the BSC tissue triplets turned out not significant. So the null hypotheses was not rejected, which means that the predicted model is probably not closer to the true model than the other models.
On the other hand, in the liver tissue, for all triplets the causality hypotheses turned out significant. So the null hypotheses was rejected, which means we can say that the predicted model – M3 in all cases – is probably closer to the true model than the other models.

Summery:

To summarize, it was observed that the liver exhibited more significant eQTLs than the blood stem cells, when it comes to our collection of SNPs. That means, there are more pairs of SNPs with genes in the liver (after the preprocessing steps on both datasets) that seem to be correlated, than in blood stem cells.
Also, when analyzing the cis and trans eQTLs we could see, as talked about is class, more significant p-values of cis-acting eQTLs than trans acting ones.

In the QTL analysis, we found more significant QTLs in the phenotype after the second injection, rather than only the first injection. That could make sense because given double the dose of the drug, the expression of the trait could be more extreme, revealing correlations we haven't seen with only one injection. We also spotted that all the eQTLs were found in a close nearby collection of SNPs on chromosome 13. That can indicate that this spot (around 80Mbp, on chromosome 13) affects cocaine response vertical activity in females.

We combined the results and found that restricting the GWAS to the DNA variants found significant with at least one gene, and found that for the most part that didn't have much effect on the eQTLs, nor led to new eQTLs. Only when limiting the GWAS to DNA variants that are found significant in at least on gene in both liver and BSC datasets (intersection), we got a new eQTL in the first phenotype that only appeared before in the second phenotype. That might be due to the nature of the FDR correction, and the small amount of eQTLs, thus relatively significant p-values to begin with.

Lastly, we performed causality analysis on the 35 triplets we found that might have a SNP-gene-trait connection (where the QTL and eQTL are in nearby genomic location).
The predicted model for all of them except 1 turned out to be M3, which means the SNP probably affects the gene and trait separately, and the gene and trait don't seem to affect each other.
We later performed a permutation test to analyze the statistical significance of results we got. On our 10 chosen triplets, only the liver tissue triplets (triplets with liver genes) showed significant p-values that led to the conclusion that the M3 model is closer to the real model than the other models.
In the BSC tissue triplets the results were non significant, which can be observed also by the much lower LR values.

We enjoyed working on this project, and finding correlations between DNA variants, gene datasets from different tissues and complex traits. We found many interesting results and correlations that might be interesting to check deeper in the future.