Metabolomics Excercise Sheet Gobi 001

Cheng Wei Lao, Alexandra Smirnova, Justin Borgmann

R Code : https://github.com/Dondonn/metabolomics

Task 4:

C4 : SNP_rs2066938

C4 = Butyrylcarnitine has the lowest p value, and therefore has the highest probability of being important for clinical or practical importance with SNP_rs2066938

Biological relevancy of finding

Butyrylcarnitine is an acyl carnitine = fatty acid with an ester bond connecting carboxylic acid to carnitine. It is related to Gastrointestinal disorders and adverse health effects, like the digestive system disorder, or Upper GI Disorder. For metabolic disorders its connected to the celiac disease. It is mainly located in the blood cells.

The SNP properties on snipa.org show for the rs2066938 that its related to one disease gene, even though the severity for this is just based on speculation and not of severe impact..yet.

The Trait annotations for rs2066938 also show the highest relation to butyrylcarnitine with the lowest p-value.

In conclusion, it seems like the gene SNP rs2066938 has some impact on human blood metabolites, like the C4, and therefore it may relate to the expression of Gastrointestinal disorders or adverse health effects connected to the digestion system, or even an autoimmune disease linked to the gastrointestinal tract like the celiac disease.

Questions of task 4:

1c:

Since we only want to focus on the relevant biological information, and reduce the influence of any measurement noise, we should get the data to a logarithmic scale

The log Transformation reduces large values in the data set relatively more than the small values, and therefore give a better scale for the whole data set to work with

1d:

When you have large datasets and do not know much about the relation of the variables, it is useful to use the three-sigma Rule(Empirical Rule) to forecast outcomes, because most data(99.7%) occurs within three standard deviations of the mean, and therefore only is relevant.

2a:

If we do not have enough information is available, it can be too random to perform statistical analysis on such data.

Linear regression is useful to predict the value of a variable based on the value of another variable, to support this and get the best prediction you can include one or more covariates, Linear regression is like a straight line that minimizes differences between predicted and actual

output values. You want to know whether one measurement variable is associated with another one, and therefore you use the linear regression.

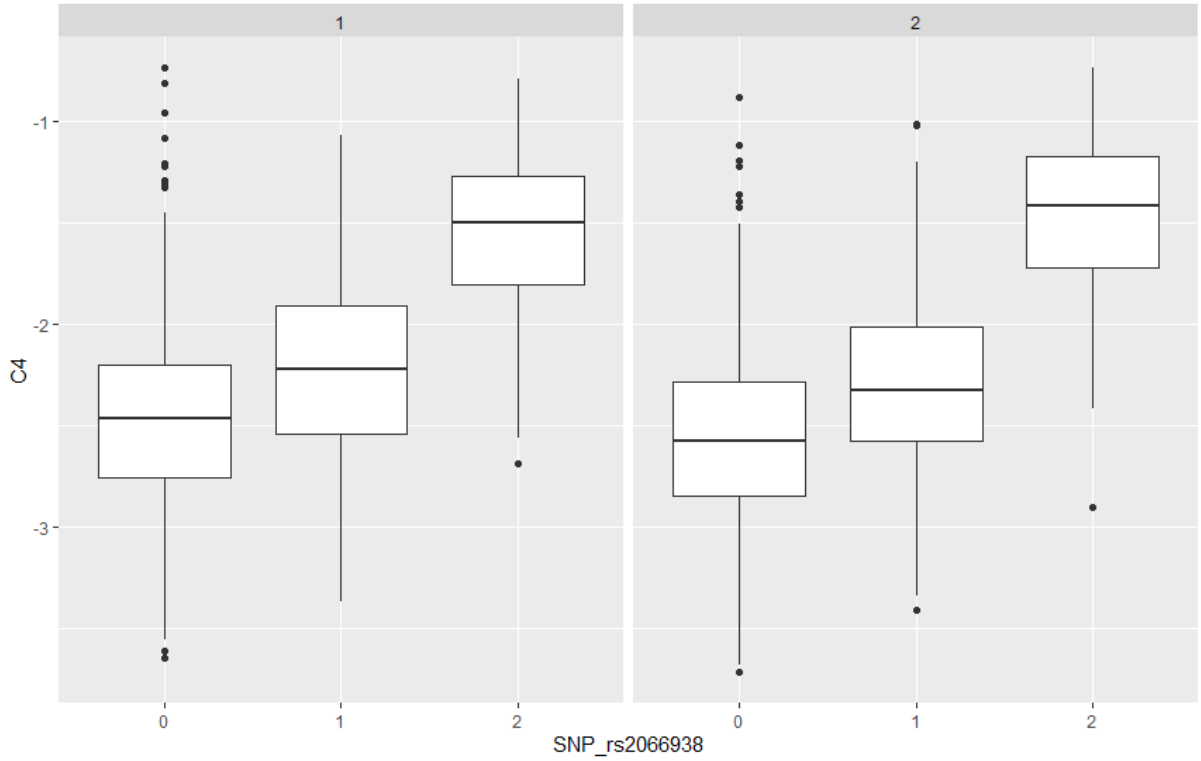
2c:

Bonferroni correction is used to compensate for Type 1 error, if you have many comparisons the type 1 error rises, and therefore you need a new threshold (different from the usual default 0.05) for the p-value, to decrease the likelihood of discovering false-positive results.

2d:

table is Significant.Associations.RData file in Github.

Boxplot:



Here is the boxplot of Butyrylcarnitine with the genotype for Men(1) and Women(2)

The male results have fewer outliers than the female results, even though there's barely a visible difference.

It seems like the concentration of the metabolite is higher in males, and compared with the biological relevance of the metabolite, and the genotype, this gives a hint of males being slightly more exposed to adverse health effects in connection to the metabolite C4.

Between the different possibilities of the genotype (0/1/2), there is a clear difference, the Homozygous genotype (2) has a higher concentration of the metabolite, and therefore is at higher risk for the negative effects the metabolite can bring with it. This makes sense, because genotype 2, or homozygous means the person has inherited the same DNA sequence of a particular gene, the person got the allele from both parents, and therefore is homozygous for that mutation in the gene.

It is also visible that for men, the 0 and 1 genotype are at a higher concentration for the metabolite than the women type 0 and 1, but for the genotype 2 of the genotype, the

concentration is actually higher.
*When talked about genotype 2, I mean the Homozygous alternate allele, since genotype 0 is the Homozygous reference allele, and has the lowest concentration of the metabolite.

Beeswarmplot:

In the beeswarmplot there is a way clearer difference between the different genotypes. While genotype 2 (Homozygous) is way less present, because its rarer, the occurrences it does have, show how the concentration is decently higher for the metabolite C4. This is the difference to the boxplot, while the boxplot does not show the genotype difference that significantly and also more the gender difference, the beeswarmplot shows the genotype difference stronger.

The homozygous genotype 0, for the reference are at the lowest risk of a high metabolite concentration, while homozygous genotype 2, for the alternate, are the highest. Genotype 1, heterozygotes is in between.