System genetics final exercise

## Submitted by: Guy Shur 311133110, Dror Hadas 302678263

# Background & Data

# Preprocessing

We have followed the preprocessing guidelines as described in the instructions. First, we have downloaded normalized data directly from the GEO FTP server. Since according to the methods section in [Mozhui et al.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3350311/) the expression data was normalized to 8 expression units without any further explanation, we have recentered the data around zero.

We then removed all metadata from the expression datasets, to be left only with expression matrices.

For each expression matrix, we’ve removed rows where the maximal expression level was below the 2.5th percentile, and rows where the variance was below the 10th percentile (-1.95, 0.02 for the liver expression dataset accordingly, and -2.91, 0.01 for the hypothalamus expression dataset accordingly)these cutoffs were determined by looking at the distribution of the max values and variance values per gene.

Since no duplicate genes were found in the expression datasets, we have ignored this guideline.

Lastly, neighboring loci were removed by comparing each row to Its subsequent row in the genotypes’ dataset, when subsequent rows with equal genotypes were found, only a single row was kept.

Preliminary to all downstream analysis, columns (i.e BXD strains) were narrowed to the intersection between each expression dataset and the genotypes dataset.

# Statistical analysis & results