Pathway Analysis of GWAS data

Solving Biological Problems that require Math 2014 Bsc2 - UNIL 30.05.2014
Alain Pulfer, Anthony Sonrel, Rosanne Miles, Stefan Milosavljevic, Supervisor: David Lamparter

Introduction

To go from SNP data to pathway scores, two steps were necessary: going from SNP data to gene scores then from gene scores to pathway scores. All the demonstrations started from the following formula:

$$Y_i = X_i b + \varepsilon_i$$

Where Y_i is the phenotype, X_i the genotype, b represents the association between genotype and phenotype and ε_i is a random variable normally distributed with mean 0 and variance σ^2. Developing the formula, the value for b that minimizes the residual error is found, scaled and a normally distributed $\hat{b}_{sc}$ is obtained. Looking at the covariance of two $\hat{b}_{sc}$, a correlation with a constant part is obtained. To obtain the correlation real data can be used.

Now that the distribution and the behaviour of the $\hat{b}_{sc}$ values are known, a gene score can be calculated. To go from gene scores to pathway scores there is a problem that should be considered; in biology, it is known that two genes that are close are often correlated and this fact can cause statistical problems. A possible solution could be considering two close genes as one (fusing them). Each fused gene is considered independent.

The next question is; should close genes be fused or not?

Conclusion

Running the first analysis, the results were that the top scoring pathways were more coherent for non-fused genes. Non-fused genes (Fig.1) also had lower p-values than fused genes (Fig.2). So at first, it seemed as though fusing the genes didn't give better results.

To make sure that the results were not false-positives, QQplots of the p-values against an uniform distribution were created (Fig.3). The graphs for non-fused genes showed more inflation than the QQplots for fused genes. This happened for both HDL results and simulations (Fig.4).

For a more precise idea of the situation, histograms were created, plotting the lowest p-value in a 100 simulations of random phenotypes (Fig.5). These histograms indicated that there were too many false-positives for non-fused genes and that the model could therefore not be trusted.

By fusing genes, the results showed less bias. The conclusion can thus be that fusing the genes ensures that results are better (that there are more true-positives) than if the genes are not fused.

Results

Figure 1

Non-fused Pathway names	Chi2 pvalue
REACTOME_LIPOPROTEIN_METABOLISM	1.33226763E-15
REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_ TRANSPORT	3.60045327E-13
REACTOME_CHYLOMICRON_MEDIATED_LIPID_TRANSPORT	2.4889979E-11
REACTOME_HDL_MEDIATED_LIPID_TRANSPORT	3.46405682E-10

Figure 2

Fused Pathway names	Chi2 pvalue
REACTOME_LIPOPROTEIN_METABOLISM	3.02701472E-7
REACTOME_HDL_MEDIATED_LIPID_TRANSPORT	1.34036602E-6
REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_TRANSPORT	3.35290932E-6

Figure 3, QQplot of HDL values

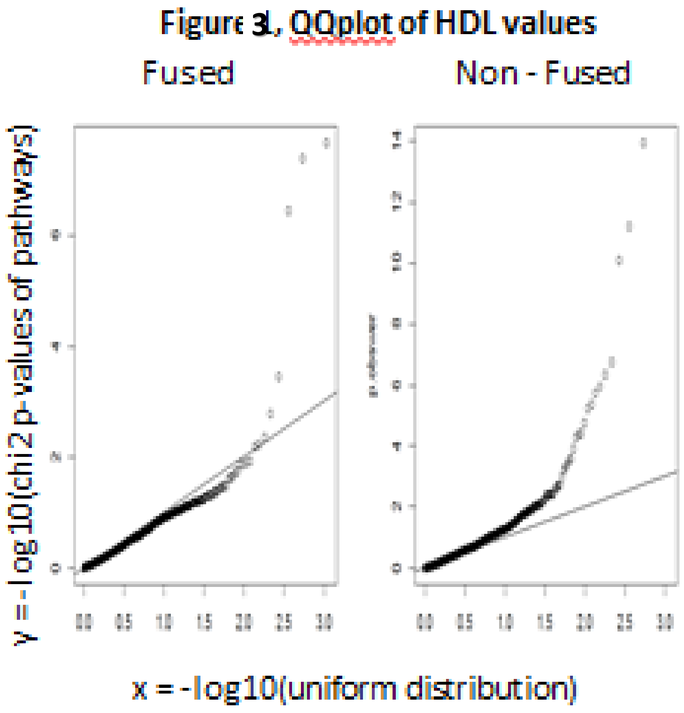

Figure 4, QQPlot of random phenotypes simulation, two different runs

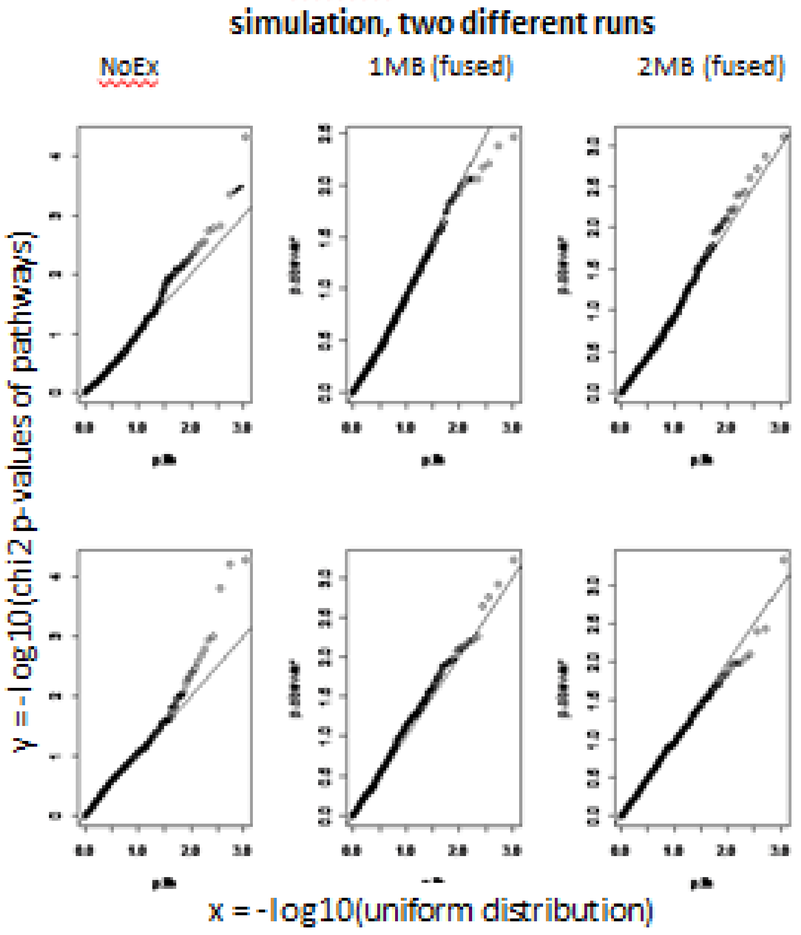

Figure 5, Histograms of p-values in 100 simulations