Exploration of Protein Phenotypes

Alvaro Barbeira

May 7, 2015

Outline

1. Problem Introduction
2. Heritability
3. mRNA Levels/Gene Expresion vs Protein
4. Bonus track: Predixcan vs Affymetrix gene expression
5. Possible Next Steps

Problem:

Genomics of Proteins

- Heritability of proteins was chosen as study subject because there are not many published results about it.
- Low data quality and noise was expected.
- Chosen data sets: hapmap release 23, Protein Data from Ron Hause and Lingfeng Wu.
- Correlations between Gene expression and Protein were also studied, using PrediXcan data.

General Method

The basic procedure was:

- Selecting individuals that satisfied $MAF >= 0.05$ with plink 1.07
- Generating Genetic Relationship Matrix with GCTA 1.24.4
- Using GCTA to figure out Restricted Maximum Likelihood (REML) for each protein

Results: Heritability took values either 0 or 1 for most genes, with standard error near 1.

Data from Hause et al results

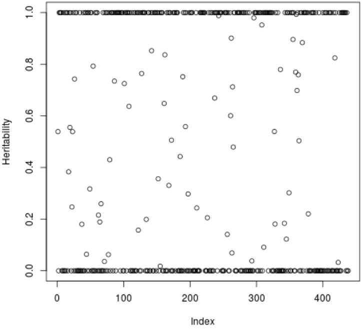

Figure : Heritability of Hause Gene Data

Data from Wu et al results

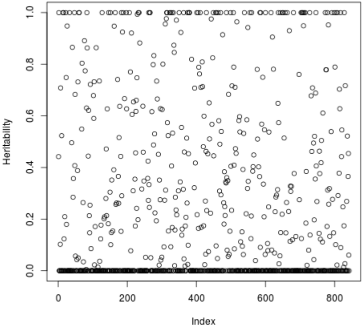

Figure : Heritability of Wu Gene Data

General Method

The procedure consisted in:

- Taking protein values for chosen people in the previous experiment.
- Figuring out correlation to mRNA levels (gene expression) data generated by PrediXcan.
- Since several proteins might correspond to a single gene, two phenotypes were studied: each protein on its own against its corresponding gene, and average of all proteins for a single gene.

Results: R^2 taken as correlation indicator was not great, but showed that some relationship could be picked up in spite of the noise. Values near 1 are taken by sets where very few individuals' data was available.

Data from Hause et al results

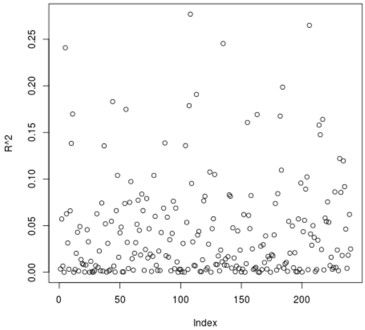

Figure : Correlation (R^2) for Hause et al. data

Data from Hause et al results

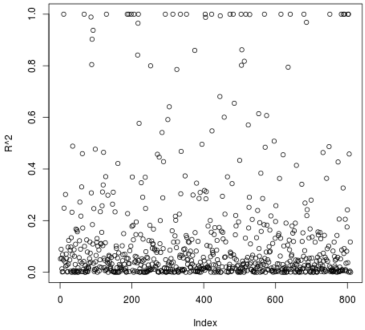

Figure : Correlation (R^2) for Wu et al. data

PrediXcan vs Affymetrix data

As a last check, predixcan data was compared to measured gene expresion from affymetrix.

- Affymetrix data had a many-to-many relationship between genes and expression measurement, so a single $(gene, measurement)$ pair was chosen for each set
- Correlation between predixcan values and affymetrix measurements were figured out

Results: Again, R^2 was not great, but showed some relationship Values near 1 are taken by sets with very few individuals' data was available.

Things to try if suddenly found idle on a sunday afternoon without anything else to do:

- Take again on heritability using protein eigenvectors as covariates in gcta
- Use PrediXcan to predict protein levels and correlate to measured proteins
- Surrogate Variable Analysis of protein data
- Figure out PCA analysis of proteins, and study heritability of residuals