UNIVERSITY OF QUEENSLAND

Epistatic control of gene expression in Humans

*** Report V1 ***

Joseph Powell
Gibran Hemani

April 21, 2013

1 INTRODUCTION

In humans expression quantitative trait loci (eQTL) have been extensively studied for their effects on transcript levels as well as their underlying effects on complex phenotypes such as disease susceptability. To date most studies have analysed SNPs independently and estimated the genotypic effect on mean expression levels assuming an additive mode of inheritance. However, epistasis (gene by gene interactions) is thought to play an important role in the control of gene expression principally through the interaction of transcription factors and the intra-cellular effects of proteins on gene regulation.

[add short paragraph outlining the types of analysis for epistasis - leading to the conclusions that 2D is the best...]

Recently we have shown that whilst most genetic variance for gene expression is likely to act in an additive manner, for many probes there exists significant non-additive genetic varaiance. [Gib: can you add something on the 'appearence of additive variance' in our system?]. The importance of epistasis for control of gene expression is largely unknown and has genrally not been studied in human populations. It is likely that this is due in part to the considerable compuation demands of running a 2D scan on 1000's of probes as well as a restrivive study-wide multiple testing burden.

[we can add a short paragraph outlining the cool work that Gib has done.]

2 METHODS

2.1 DISCOVERY DATA

The Brisbane Systems Genetics Study (BSGS) comprises 862 Individuals of European descent from 274 independent families [2]. DNA samples from each individual were genotyped on the Illumina 610-Quad Beadchip by the Scientific Services Division at deCODE Genetics Iceland. Full details of genotyping procedures are given in Medland et al. [1] Standard QC filters were applied and the remaining 528,509 SNPs were carried forward further analysis.

Gene expression profiles were generated from whole blood collected with PAXgene TM tubes (QIAGEN, Valencia, CA) using Illumina HT12-v4.0 bead arrays. The Illumina HT-12 v4.0 chip contains 47,323 probes, although some probes are not assigned to RefSeq genes. We removed any probes that did not match the following criteria: contained a SNP within the probe sequence with MAF > 0.05 within 1000 genomes data; did not map to a listed ref-seq gene; were not significantly expressed (based on a detection p-value < 0.05) in 100% of samples. After this stringent QC 7000 [need to get exact number] probes remained for 2d-eqtl mapping.

2.2 NORMALISATION

Gene expression profiles were normalised and batch effects removed with the following procedures; Raw background expression were removed for each sample. Expression levels were then adjusted using Quantile and log2 transformation to standardise distribution between samples.To avoid inflation of the test statistic due to polygenic effects the following linear model was used to correct for batch and polygeneic effects;

$$y = \mu + c + p + s + a + g + e \tag{2.1}$$

where μ is the population mean expression levels, c, p, s and a are vectors of chip, chip position, sex and generation respectivly. g is a random additive polygenic effect with a variance covariance matrix

$$G_{ijk} = \begin{cases} \sigma_a^2 + \sigma_e^2 & j = k \\ 2\phi_{ijk}\sigma_a^2 & j \neq k \end{cases} \tag{2.2}$$

The parameters σ_a^a and σ_e^2 are variance components for additive background genetic and environmental effects respectively. Here, we are using family based pedigree information rather than SNP based IDB to account for relationships between individuals and so ϕ_{ijk} is the kinship coefficient between individuals j and k. The residual, e, from equation 2.1 are assumed to follow a multivariate normal distribution with a mean of zero. These residuals are corrected for batch, sex, generation and polygentic effects. Residuals were standardised to Z-scores uning a inverse-normal transformation and used as the adjusted phenotype for the pairise epistasis scan.

2.3 EXHAUSTIVE 2D-EQTL ANALYSIS

We used epi-GPU software to perform an exhaustive scan for pairwise interactions, such that each SNP is tested against all other SNPs for statistical association with the expression values

for each of the 7000 probes. For each SNP-pair there are 9 possible genotype classes. We treat each genotype class as a fixed effect and fit an 8 d.f. F-test to test the following hypotheses:

$$H_0 : \sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{x}_{ij} - \mu)^2 = 0; \tag{2.3}$$

$$H_1 : \sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{x}_{ij} - \mu)^2 > 0; \tag{2.4}$$

where μ is the mean expression level and x_{ij} is the pairwise genotype class mean for genotype i at SNP 1 and genotype j at locus B. This type of test does not parameterisze for specific types of epistasis, rather it tests for the joint genetic effects at two loci. This has been demonstrated to be statistically more efficient thwhen searching for a wide range of epistatic patterns (add refs), although will include any marginal effects of SNPs.

The complete exhaustive scan for 7,000 probes comprises of ~$9.8e^{14}$ F-tests. Due to the very large number of tests [plus nature of epistasis ....] we used a highly conservative series of filtering steps to identify study-wide significant epistasis associations to follow forward for replication analysis. Initially a significance threshold of $-log10$ 15.5. This threshold is the studywide significance threshold determined by the number of independent tests for a 2D scan for a single probe / the number of probes analysed. This is likely to be conservative as this threshold assumes independance between probes. Following filtering on this threshold only SNP pairs with all 9 genotype classes and minium genotype class size > 5 individuals were kept. We then calculated the linkage disequilibrium between SNPs in a pair and removed any pairs with $r^2 > 0.1$. We also removed any pairs containing SNPs for which single marker additive or dominace effects had been identifyed in previous work (Powell et al. in Press). Sentinal SNP pairs we retained from epistasis eQTL 'peaks' comprising of multiple sets of pair-wise SNPs. At this stage ~13,000 SNP pairs remained and were tested for the contribution of marginal and interaction terms by testing the 8d.f. 2.4 verses a 4 d.f. model parameterizing for interaction terms only:

$$H_0 : \sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \mu)^2 = 0; \tag{2.5}$$

$$H_1 : \sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \mu)^2 > 0; \tag{2.6}$$

where $\bar{x}_i(\bar{x}_j)$ is the marginal class mean for genotype $i(j)$ at SNP A (B). Significance of interaction terms was determined using a bonferroni threshold of $0.05/13000 = 3.8e^{-6}$

2.4 REPLICATION

2.5 SIMULATION

REFERENCES

[1] Sarah E Medland, Dale R Nyholt, Jodie N Painter, Brian P McEvoy, Allan F McRae, Gu Zhu, Scott D Gordon, Manuel a R Ferreira, Margaret J Wright, Anjali K Henders, Megan J Campbell, David L Duffy, Narelle K Hansell, Stuart Macgregor, Wendy S Slutske, Andrew C Heath, Grant W Montgomery, and Nicholas G Martin. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American journal of human genetics*, 85(5):750–5, November 2009.

[2] J E Powell, A K Henders, A F McRae, A Caracella, S Smith, M J Wright, J B Whitfield, E T Dermitzakis, N G Martin, P M Visscher, and G W Montgomery. The brisbane systems genetics study: genetical genomics meets complex trait genetics. *PLoS ONE*, 7(4):e35430, 2012.