UNIVERSITY OF QUEENSLAND

# Epistasis follow-up

## *** summary ***

# ORIGINAL ANALYSIS SUMMARY

Original discovery analysis identified 501 interactions comprising of 781 unique SNPs and 238 genes (probes). These were 26 cis-cis, 462 cis-trans and 13 trans trans. The majority of our discovery interactions were composed of one SNP that was significantly associated with the gene expression level in the discovery data set, and one SNP that had no previous association (439 out of 501, Methods). Only nine interactions were between SNPs that both had known main effects, whereas 64 were between SNPs that had no known main effects.

In the original analysis the following thresholds were used;

## STAGE 1

The complete exhaustive scan for 7339 probes comprises $1.03 \times 10^{15}$ $F$-tests. We used permutation analysis to estimate an appropriate significance threshold for the study. To do this we performed a further 1600 exhaustive 2D scans on permuted phenotypes to generate a null distribution of the extreme $p$-values expected to be obtained from this number of multiple tests given the correlation structure between the SNPs. We took the most extreme $p$-value from each of the 1600 scans and set the 5% FWER to be the 95% most extreme of these $p$-values, $T_* = 2.13 \times 10^{-12}$. The effective number of tests in one 2D scan being performed is therefore $N_* = 0.05/T_* \approx 2.33 \times 10^{10}$. To correct for the testing of multiple probes we established an experiment wide threshold of $T_e = 0.05/(N_* \times 7339) = 2.91 \times 10^{-16}$.

FILTERING We used two approaches to filter SNPs from stage 1 to be tested for significant interaction effects in stage 2.

FILTER 1 After keeping SNP pairs that surpassed the $2.91 \times 10^{-16}$ threshold in stage 1 only SNP pairs with at least 5 data points in all 9 genotype classes were kept. We then calculated the LD between interacting SNPs (amongst unrelated individuals within the discovery sample and also from 1000 genomes data) and removed any pairs with $r^2 > 0.1$ or $D'^2 > 0.1$ to avoid the inclusion of haplotype effects and to increase the accuracy of genetic variance decomposition. If multiple SNP pairs were present on the same chromosomes for a particular expression trait then only the sentinel SNP pair was retained, *i.e.* if a probe had multiple SNP pairs that were on chromosomes one and two then only the SNP pair with the most significant $p$-value was retained. At this stage 6404 filtered SNP pairs remained.

FILTER 2 We also performed a second filtering screen applied to the list of SNP pairs from stage 1 that was identical to filter 1 but an additional step was included where any SNPs that

had previously been shown to have a significant additive or dominant effect ($p < 1.29 \times 10^{-11}$) were removed, creating a second set of 4751 unique filtered SNP pairs.

## STAGE 2

To ensure that interacting SNPs were driven by epistasis and not marginal effects we performed a nested ANOVA on each pair in the filtered set to test if the interaction terms were significant. We did this by contrasting the full genetic model (8 *d.f.*) against the reduced marginal effects model which included the additive and dominance terms at both SNPs (4 *d.f.*). Thus, a 4 *d.f.* $F$-test was performed on the residual genetic variation, representing the contribution of epistatic variance. Significance of epistasis was determined using a Bonferroni threshold of $0.05/(6404 + 4751) = 4.48 \times 10^{-6}$. This resulted in 406 and 95 SNP pairs with significant interaction terms from filters 1 and 2, respectively.

## TYPE 1 ERROR RATE

Using a Bonferroni correction of 0.05 in the second stage of the two stage discovery scan implies a type 1 error rate of $\alpha = 0.05$. However, this could be underestimated because the number tests performed in the second stage depends on the number of tests in the first stage, and this depends on statistical power and model choice. We performed simulations to estimate the type 1 error rate of this study design.

We assumed a null model where there was one true additive effect and 7 other terms with no effect. To simulate a test statistic we simulated 8 z-scores, $z_1 \sim N(\sqrt{NCP}, 1)$ and $z_{2..8} \sim N(0, 1)$. Thus $z_{full} = \sum_{i=1}^{8} z_i \sim \chi^2_8$ (representing the 8 d.f. test) and $z_{int} = \sum_{i=5}^{8} z_i \sim \chi^2_4$ (representing the 4 d.f. test where the null hypothesis of no epistasis is true). For a particular value of $NCP$ we simulated 100,000 $z$ values, and calculated the $p_{full}$-value for the $z_{full}$ test statistic. The $n_{int}$ test statistics with $p_{full} < 2.31 \times 10^{-16}$ were kept for the second stage, where the type 1 error rate of stage 2 was calculated as the proportion of $p_{int} < 0.05/n_{int}$. The power at stage 1 was calculated as $n_{int}/100,000$. This procedure was performed for a range of $NCP$ parameters that represented power ranging from $\sim 0$ to $\sim 1$.

## METHODS AND RESULTS

The following analyses have been conducted;

1. Determining the empirical p-values for each of the 501 interactions

The initial analysis used F-tests and some simulation work to determine the expected Type 1 error rate in the 1st stage of the discovery process. The 1st stage was followed by a 2nd stage where the interaction model was fitted. Subsequent simulations and theoretical calculations have suggested that the Type 1 error rate of the 2nd stage is not correct when there is a large main effect and / or in the presence of LD.

a. Identify which of the two snps in the original epistasis pair has the largest additive effect.

b. Treat the largest additive SNP as a fixed SNP and perform a genome-wide analysis using the 8df and 4df epistasis model.

c. This generates $\approx 500,000$ interaction p-values. Apply the same snp-pair filtering as used in the manuscript. Namely, LD ($r^2 < 0.1$), nclass = 9, and minclass > 5. Any SNP with +/- 5MB of original epistasis SNP pairs were also removed.

d. Use the (filtered) interaction p-values to determine the empirical distribution of null p-values.

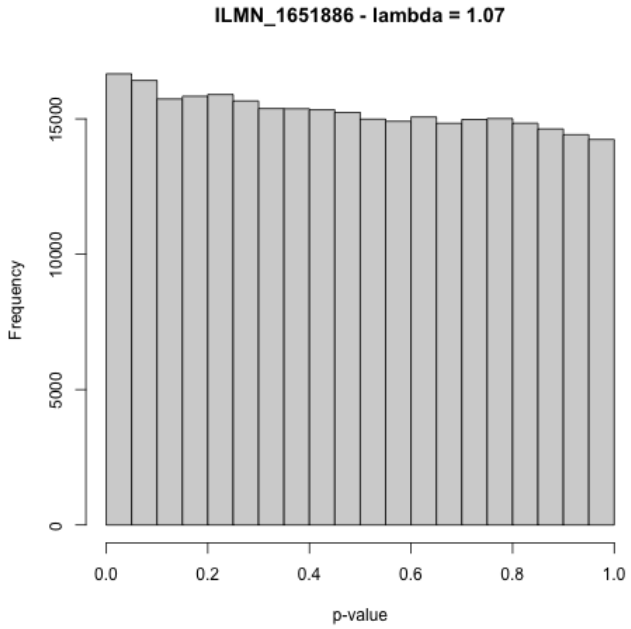


Figure | p-value distribution - ok

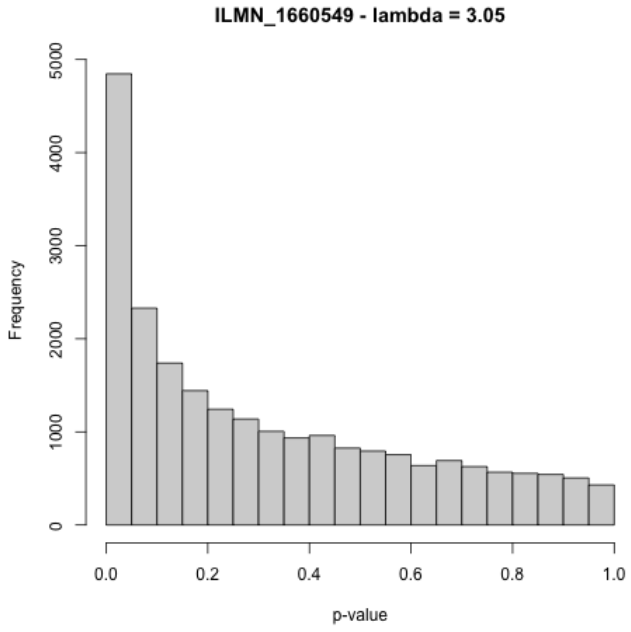


Figure | p-value distribution - bad

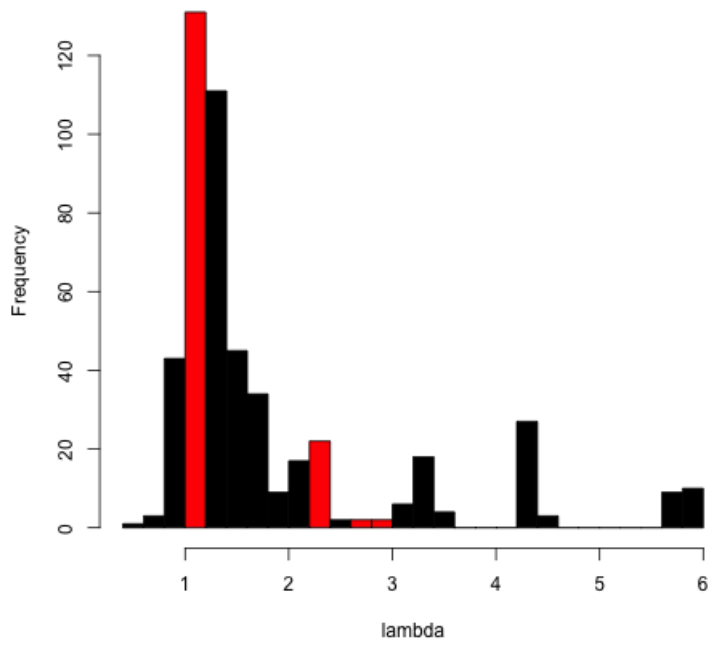


Figure | Distribution of median lambda from the 501 fixed SNP analysis. Red denotes original pairs identified from filter 2 - i.e. no previous study-wide additive effects. Black denotes the original filter 1 pairs

2. Identify the largest additive eQTL for the probe (irrespective of effect size). Regress out the effect of the additive loci and use the adjusted phenotype for 8df and 4df model analysis. Results reported are the 8df and 4df of original pair and the empirical p-values from genome-wide analysis fitting each of the two SNPs as fixed. As before these are determined by filtering out the SNPs on the same chromosome as the original fixed SNP and also within +/- 5MB of the second SNP.

[Currently running]

2. Prediction

Of the 501 SNP pairs, 484 have both SNP in the EGCUT data. Most of the Inchianti SNPs need to be imputed, but we expect most to pass filtering. For pairs without and Inchianti SNP I propose using the SNP with the largest additive effect in the egcut data.

For each pair;

a. Predict the phenotype in egcut data using a predictor with effects estimated from

4df model (estimated in BSGS)
8df model (estimated in BSGS)
1df model (estimated in BSGS using the Inchianti snp)