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 1.03×10^{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 2.91×10^{-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 ~ 0 to ~ 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.

	F	P	nclass	minclass	LD
1	0.36	0.78	8.00	1.00	0.00
2	0.44	0.72	8.00	7.00	0.00
3	0.33	0.85	9.00	2.00	0.00
4	0.36	0.84	9.00	6.00	0.01
5	0.29	0.75	6.00	1.00	0.00
6	0.12	0.89	6.00	1.00	0.00
...	...	...	...	...	...

Header of output for a single pair

probename	snp1	snp2	nclass9	minclass5	LD01	npass	nsnps	filter	p_egcut	p_fehr	λ	nthres	Fe	N_Fe	F_emp	P_emp	type1
ILMN_1651385	rs7989895	rs4846085	419387	285750	506818	269121	506818	1.00	0.25	0.61	1.61	4	9.43	0	8.95	6.35	0.10
ILMN_1651705	rs872311	rs11032695	419021	278721	511121	260851	511121	1.00	0.30	0.26	1.12	1	9.41	0	7.75	5.41	0.05
ILMN_1651886	rs7108734	rs12784396	427787	319167	501291	305496	501291	1.00	0.01	0.21	1.07	0	9.5	0	7.4	5.13	0.05
ILMN_1652333	rs898095	rs9892064	442554	336649	515007	322786	515007	2.00	29.39	28.24	1.34	7	9.53	0	9.48	6.77	0.07
ILMN_1653205	rs12429804	rs2896452	383323	213815	507099	183628	507099	1.00	0.02	0.29	1.78	24	9.22	4	10.56	7.63	0.11
ILMN_1653205	rs12454561	rs2896452	386460	215536	511390	185098	511390	1.00		0.31	1.78	25	9.22	4	10.56	7.63	0.11
ILMN_1653205	rs2896452	rs1004564	373839	208706	494954	179170	494954	1.00	0.18	0.38	1.78	23	9.2	2	10.08	7.25	0.11
ILMN_1653205	rs7152284	rs2896452	385471	215221	509793	184859	509793	1.00	0.07	2.18	1.78	23	9.22	3	10.56	7.63	0.11
ILMN_1653205	rs8051751	rs2896452	386341	215644	511501	185188	511501	1.00	0.18	1.39	1.78	24	9.22	4	10.56	7.63	0.11
ILMN_1654545	rs4333645	rs1455268	421112	307876	495374	294544	495374	2.00	0.01	0.10	1.45	4	9.48	0	9.14	6.5	0.08
ILMN_1656378	rs10906857	rs12490878	418379	314668	489831	301252	489831	1.00	0.34	0.42	1.1	1	9.49	0	7.69	5.36	0.05
ILMN_1658247	rs11613438	rs1047944	402228	251785	490871	234981	490871	2.00	1.55	1.27	1.53	21	9.35	1	9.73	6.97	0.09

Header of summary output for 501 pairs

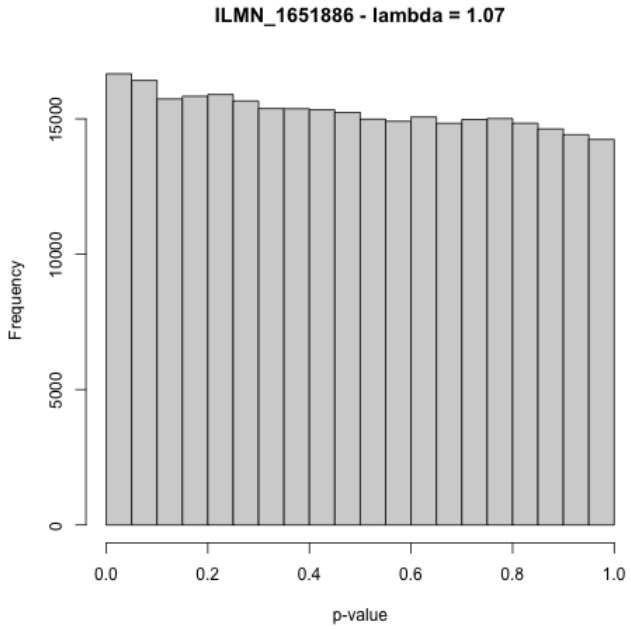

Figure | Example of p-value distribution - ok

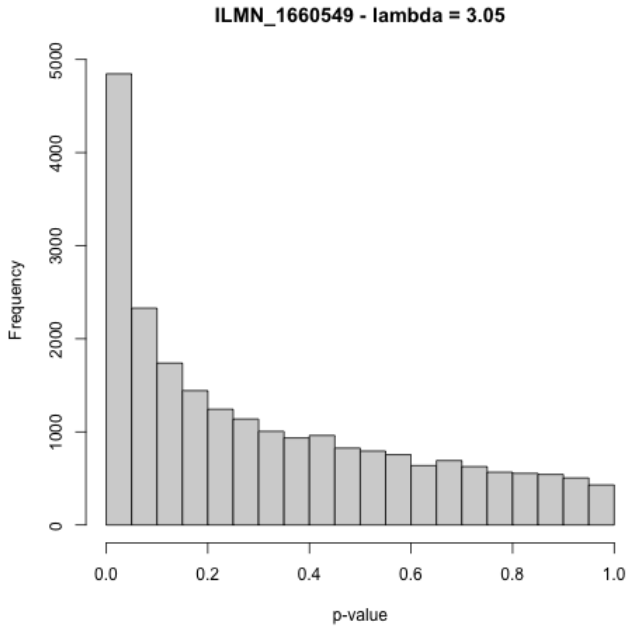

Figure | Example of p-value distribution - bad

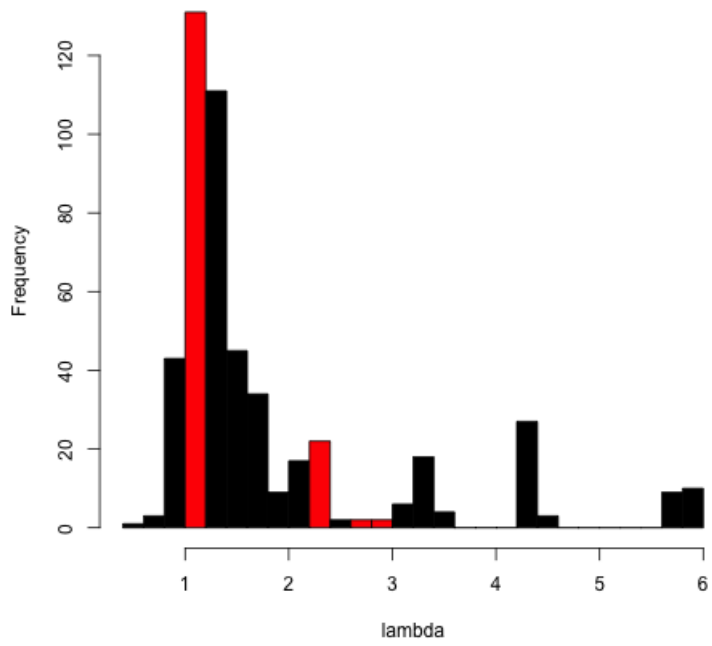

Figure | Distribution of median lambda from the 501 fixed SNP analysis. Red denotes original pairs identified from filter 2 - i.e. no previous study-wide additive effects. Black denotes the original filter 1 pairs

probename	snp1	snp2	meanlambda	npairs
ILMN_1704730	rs1884655	rs10255470	2.88	10
ILMN_1710752	rs2123758	rs2786014	2.15	8
ILMN_1717234	rs1157079	rs7733671	4.31	17
ILMN_1720059	rs12435486	rs7837237	2.29	7
ILMN_1738784	rs10930170	rs12120009	2.24	6
ILMN_1755589	rs11080134	rs11169322	1.16	6
ILMN_1786426	rs2839013	rs8106959	5.65	20
ILMN_1804396	rs1293455	rs2655991	1.38	7
ILMN_2313158	rs10869600	rs13069559	3.15	15
ILMN_2372639	rs17159840	rs10059004	4.17	17
ILMN_3231952	rs12947580	rs8079215	2.16	6

Mean lambda per probe, when 5 or more interactions are identified amongst the 501 pairs

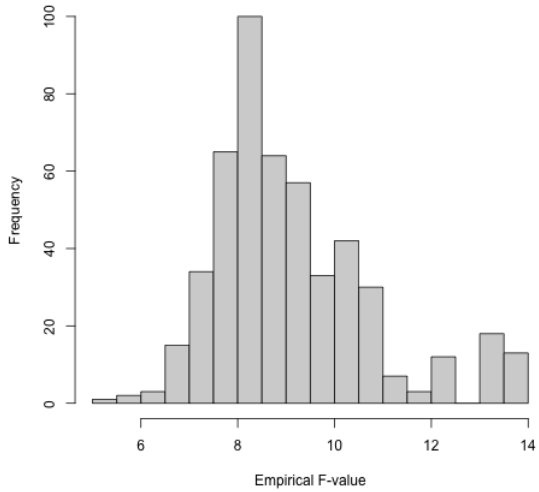

Figure | The F-statistic corresponding to $p = 4.48^{-6}$ and df1=4, df2=842 from a H_0 table is 7.67. This figure shows the empirical (ranked) F-statistic corresponding to $n * 4.48^{-6}$. Where $n * 4.48^{-6} < 1$, the largest F-statistic taken.

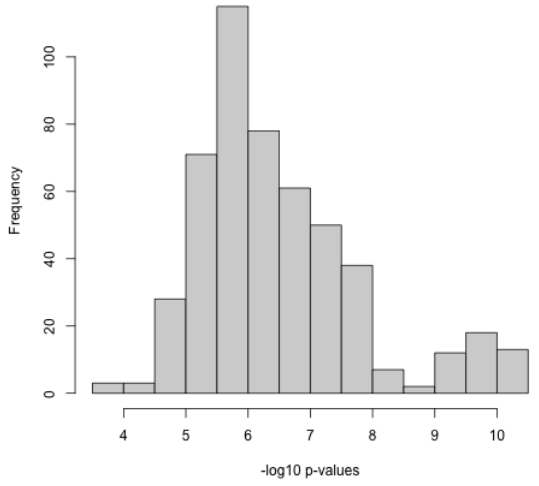

Figure | The -log10 p-values corresponding to above F-statistics (df1=4, df2=842). The -log10 $p = 4.48^{-6} = 5.35$

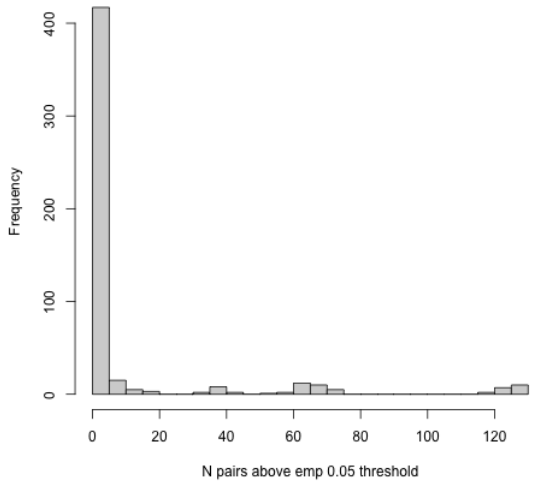

Figure | For each pair (n=501) the number of tests where the test statistic is greater than the 95th percentile of an F-dist

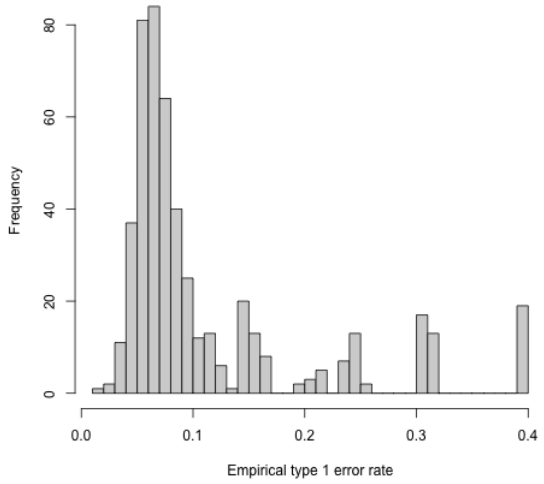

Figure | Distribution of the empirical type 1 error rate. Determined as the proportion of tests with $F > 2.38$ (95th percentile of F-dist with df1=4, df2=842)

gene	probename	snp1	snp2	λ	Type1
ADK	ILMN_2358626	rs2395095	rs10824092	1.15	0.06
ATP13A1	ILMN_2134224	rs4284750	rs873870	1.82	0.119
C21ORF57	ILMN_1795836	rs9978658	rs11701361	1.43	0.085
CSTB	ILMN_1761797	rs9979356	rs3761385	1.05	0.053
CTSC	ILMN_2242463	rs7930237	rs556895	1.12	0.054
FN3KRP	ILMN_1652333	rs898095	rs9892064	1.34	0.075
GAA	ILMN_2410783	rs11150847	rs12602462	1.38	0.081
HNRPH1	ILMN_2101920	rs6894268	rs4700810	1.47	0.091
LAX1	ILMN_1769782	rs1891432	rs10900520	1.22	0.066
MBLN1	ILMN_2313158	rs11981513	rs13069559	3.36	0.242
MBLN1	ILMN_2313158	rs16864367	rs13079208	2.28	0.161
MBLN1	ILMN_2313158	rs2030926	rs13069559	3.39	0.241
MBLN1	ILMN_2313158	rs218671	rs13069559	3.36	0.242
MBLN1	ILMN_2313158	rs2614467	rs13069559	3.37	0.242
MBLN1	ILMN_2313158	rs7710738	rs13069559	3.38	0.243
MBP	ILMN_2398939	rs8092433	rs4890876	1.19	0.066
NAPRT1	ILMN_1710752	rs2123758	rs3889129	2.14	0.155
NCL	ILMN_2121437	rs7563453	rs4973397	1.2	0.065
PRMT2	ILMN_1675038	rs2839372	rs11701058	2.77	0.199
SNORD14A	ILMN_1799381	rs2634462	rs6486334	2.45	0.168
TMEM149	ILMN_1786426	rs807491	rs7254601	2.69	0.194
TMEM149	ILMN_1786426	rs8106959	rs1843357	5.79	0.396
TMEM149	ILMN_1786426	rs8106959	rs2351458	5.77	0.395
TMEM149	ILMN_1786426	rs8106959	rs6718480	5.82	0.397
TMEM149	ILMN_1786426	rs8106959	rs6926382	5.75	0.396
TMEM149	ILMN_1786426	rs8106959	rs914940	5.76	0.395
TMEM149	ILMN_1786426	rs8106959	rs9509428	5.82	0.398
VASP	ILMN_1743646	rs1264226	rs2276470	1.37	0.08
RPL13	ILMN_2413278	rs352935	rs2965817	1.11	0.058
TRA2A	ILMN_1731043	rs7776572	rs11770192	1.36	0.079

The λ and type 1 error rates for the 30 (bonferroni-correction) significant replicated pairs mentioned in Hemani *et al.*

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)