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 $z_{full} = \sum_{i=1}^{8} z_i \sim \chi_8^2$ (representing the 8 d.f. test) and $z_{int} = \sum_{i=5}^{8} z_i \sim \chi_4^2$ (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. GWAS based determination of 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. The SNP with largest additive effect was identified for each pair of the 501 original epistasis SNPs.

b. The largest additive SNP was treated as a fixed SNP and a genome-wide analysis using the 8df and 4df epistasis model was performed.

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

d. The filtered interaction p-values were used to determine the empirical distribution of null p-values.

e. Summary information such as median lambda were calculated from the filtered interaction p-values.

2. Permutation based determination of the empirical p-values for each of the 501 interactions

a. As before, the SNP with largest additive effect was identified for each pair of the 501 original epistasis SNPs.

b. Genotypes at the corresponding epistatic "co-SNP" were randomly shuffled (no replacement) amongst individuals.

c. Interaction p-values and text statistics were calculated using the same methods.

d. For each pair (n=501) this analysis was performed 10,000,000 times.

e. The same filtering was applied (although r^2 not required).

f. The empirical p-value determined based on rank of observed F-Statistic

	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	rs8722311	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
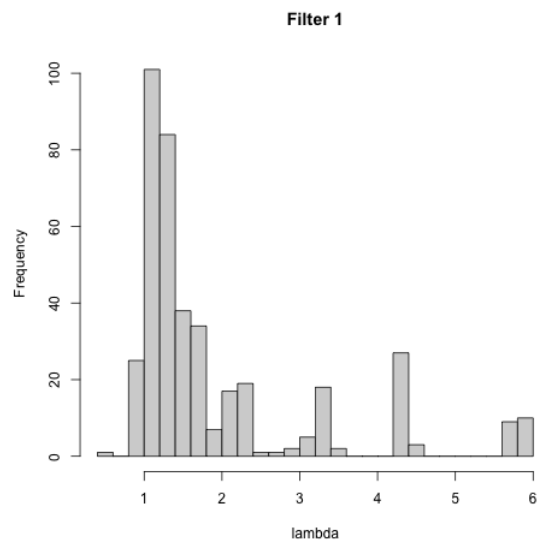
Figure | Distribution of median lambda from the 406 filter 1 pairs from the GWAS based analysis. The mean lambda is 1.94
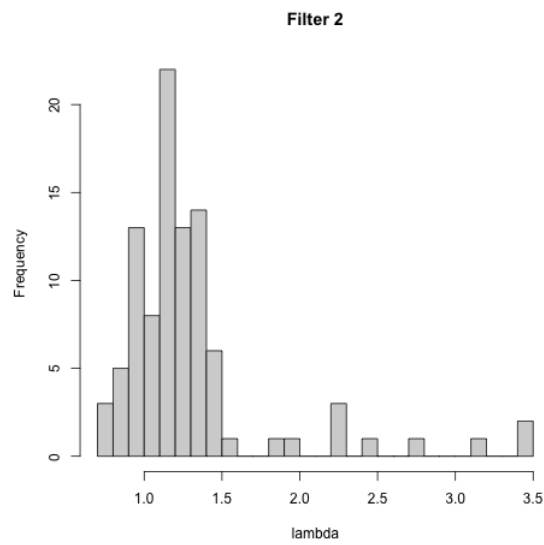

Figure | Distribution of median lambda from the 95 filter 2 pairs from the GWAS based analysis. The mean lambda is 1.31

probename	snp1	snp2	gene	meanlambda	npairs
ILMN_1704730	rs1884655	rs10255470	CD93	2.88	10
ILMN_1710752	rs2123758	rs2786014	NAPRT1	2.15	8
ILMN_1717234	rs1157079	rs7733671	CAST	4.31	17
ILMN_1720059	rs12435486	rs7837237	HMBOX1	2.29	7
ILMN_1738784	rs10930170	rs12120009	PPP2R5A	2.24	6
ILMN_1755589	rs11080134	rs11169322	DIP2B	1.16	6
ILMN_1786426	rs2839013	rs8106959	TMEM149	5.65	20
ILMN_1804396	rs1293455	rs2655991	C14ORF4	1.38	7
ILMN_2313158	rs10869600	rs13069559	MBNL1	3.15	15
ILMN_2372639	rs17159840	rs10059004	TRAPPC5	4.17	17
ILMN_3231952	rs12947580	rs8079215	ARL17B	2.16	6

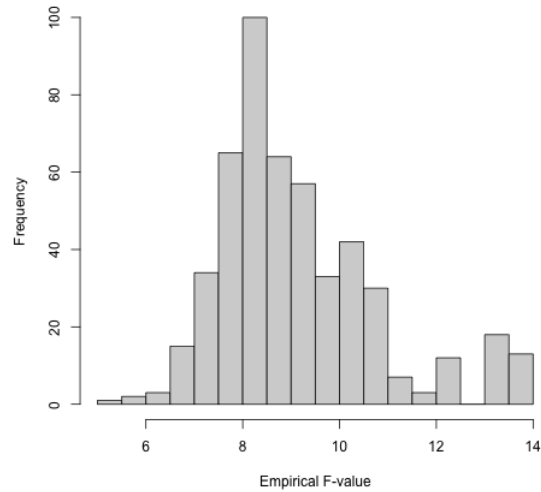This table shows the mean lambda per probe, when 5 or more interactions are identified amongst the 501 pairs

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 is taken. There are 96 pairs where the corresponding F statistic is less than 7.67
. None of these are amongst the 30 Bonferroni correction significant pairs.
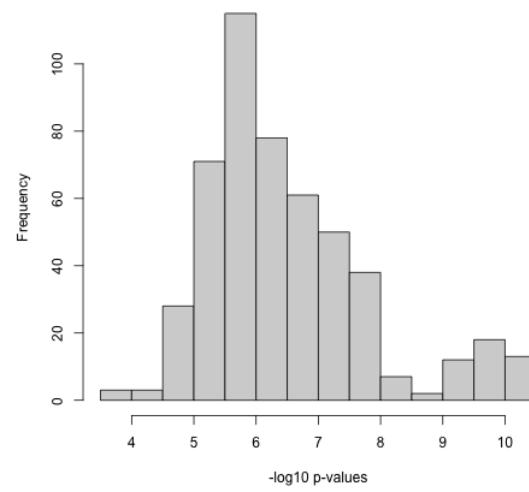
Figure | The -log10 p-values corresponding to above F-statistics (df1=4, df2=842). The -log10 $p = 4.48^{-6} = 5.35$
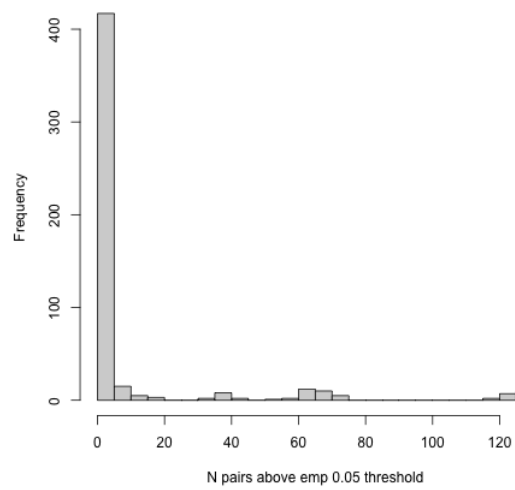
Figure | For each pair (n=501) the number of tests where the test statistic is greater than the 95th percentile of an F-dist
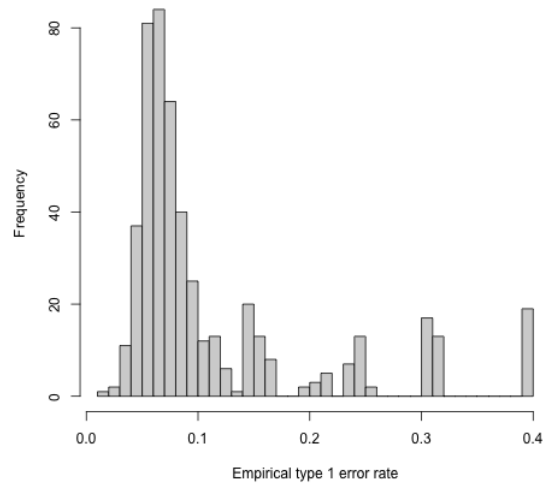

Figure | Distribution of the empirical type 1 error rate. Determined as the proportion of tests with $F > 2.38$ (95th percentile of F-dist with df1=4, df2=842)
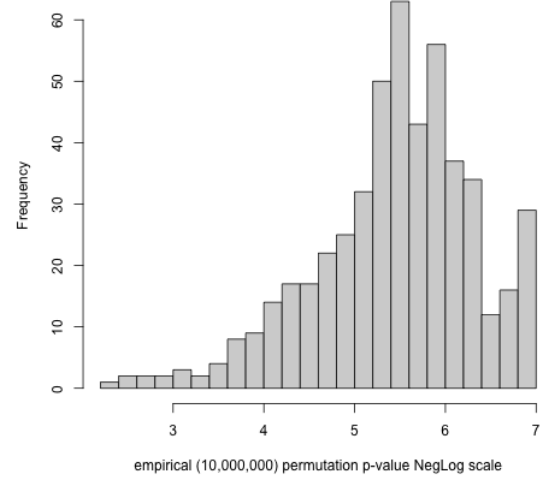
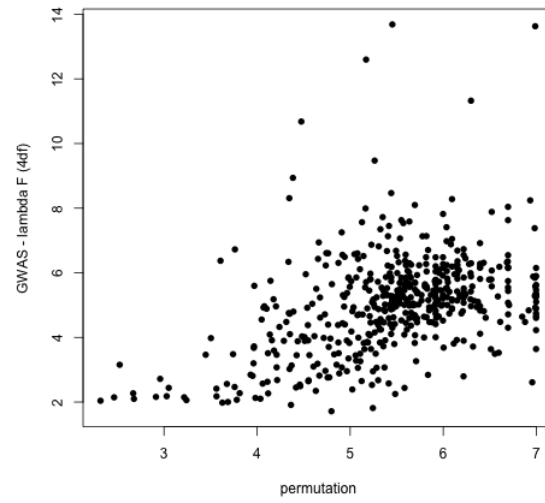Figure | Distribution of the permutation empirical p-values

Figure | the relationship between the observed p values corrected for lambdaF, which was calculated from the GWAS-based analysis and the permutation empitical p-values
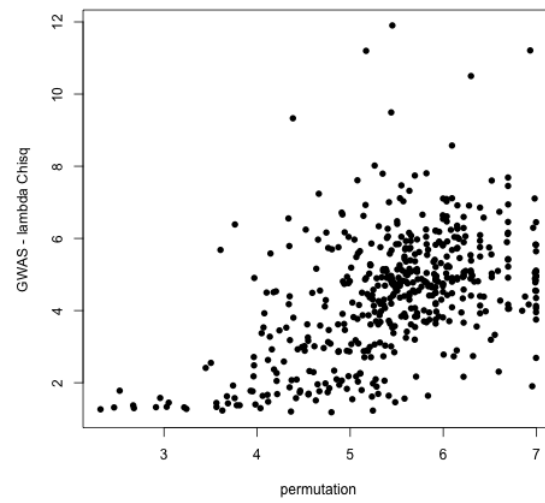

Figure | the relationship between the observed p values corrected for lambdaChisq, which was calculated from the GWAS-based analysis and the permutation empitical p-values
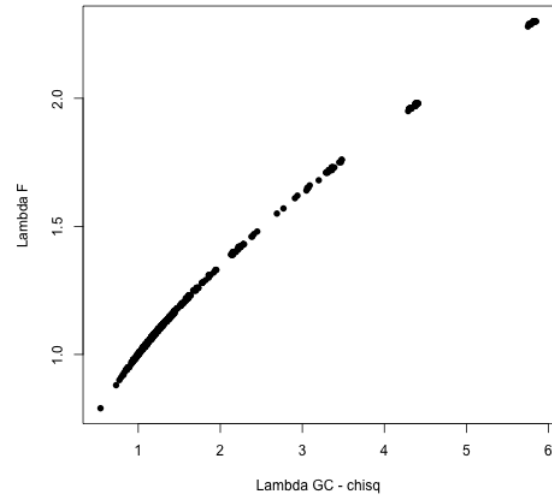
Figure | the relationship between the lambdaChisq and lambdaF from the GWAS-based analysis

gene	probename	snp1	snp2	filter	gene.1	nthreshold	P_emp
ADK	ILMN_2358626	rs2395095	rs10824092	1	ADK	3	7.74
ATP13A1	ILMN_2134224	rs4284750	rs873870	2	ATP13A1	23	7.08
C21ORF57	ILMN_1795836	rs9978658	rs11701361	2	C21ORF57	6	5.65
CSTB	ILMN_1761797	rs9979356	rs3761385	2	CSTB	4	7.51
CTSC	ILMN_2242463	rs7930237	rs556895	1	CTSC	1	5.59
FN3KRP	ILMN_1652333	rs898095	rs9892064	2	FN3KRP	7	6.77
GAA	ILMN_2410783	rs11150847	rs12602462	2	GAA	9	6.91
HNRPH1	ILMN_2101920	rs6894268	rs4700810	2	HNRPH1	13	7.14
LAX1	ILMN_1769782	rs1891432	rs10900520	2	LAX1	5	6.64
MBLN1	ILMN_2313158	rs11981513	rs13069559	1	MBNL1	86	9.06
MBLN1	ILMN_2313158	rs16864367	rs13079208	2	MBNL1	79	7.22
MBLN1	ILMN_2313158	rs2030926	rs13069559	1	MBNL1	91	9.06
MBLN1	ILMN_2313158	rs218671	rs13069559	1	MBNL1	98	9.06
MBLN1	ILMN_2313158	rs2614467	rs13069559	1	MBNL1	95	9.06
MBLN1	ILMN_2313158	rs7710738	rs13069559	1	MBNL1	89	9.06
MBP	ILMN_2398939	rs8092433	rs4890876	2	MBP	3	5.9
NAPRT1	ILMN_1710752	rs2123758	rs3889129	1	NAPRT1	80	7.36
NCL	ILMN_2121437	rs7563453	rs4973397	2	NCL	4	5.88
PRMT2	ILMN_1675038	rs2839372	rs11701058	2	PRMT2	39	7.07
SNORD14A	ILMN_1799381	rs2634462	rs6486334	2	SNORD14A	63	7.22
TMEM149	ILMN_1786426	rs807491	rs7254601	1	TMEM149	74	7.4
TMEM149	ILMN_1786426	rs8106959	rs1843357	1	TMEM149	148	9.71
TMEM149	ILMN_1786426	rs8106959	rs2351458	1	TMEM149	154	9.71
TMEM149	ILMN_1786426	rs8106959	rs6718480	1	TMEM149	146	9.71
TMEM149	ILMN_1786426	rs8106959	rs6926382	1	TMEM149	157	9.71
TMEM149	ILMN_1786426	rs8106959	rs914940	1	TMEM149	156	9.71
TMEM149	ILMN_1786426	rs8106959	rs9509428	1	TMEM149	160	9.71
VASP	ILMN_1743646	rs1264226	rs2276470	2	VASP	10	6.81
RPL13	ILMN_2413278	rs352935	rs2965817	2	RPL13	2	6.57
TRA2A	ILMN_1731043	rs7776572	rs11770192	2	TRA2A	7	6.84

The λ and type 1 error rates for the 30 (Bonferroni-correction) significant replicated pairs mentioned in Hemani *et al.* P_emp demotes the p-value from the observed F-statistic (df1=4, df2=846) which corresponds to $n * 4.44e^{-6}$ ranked test statistic from the âĂŸgwasâĂŹ analysis. If $n * 4.4e^{-6} < 1$ then the top ranked F-stat was used.
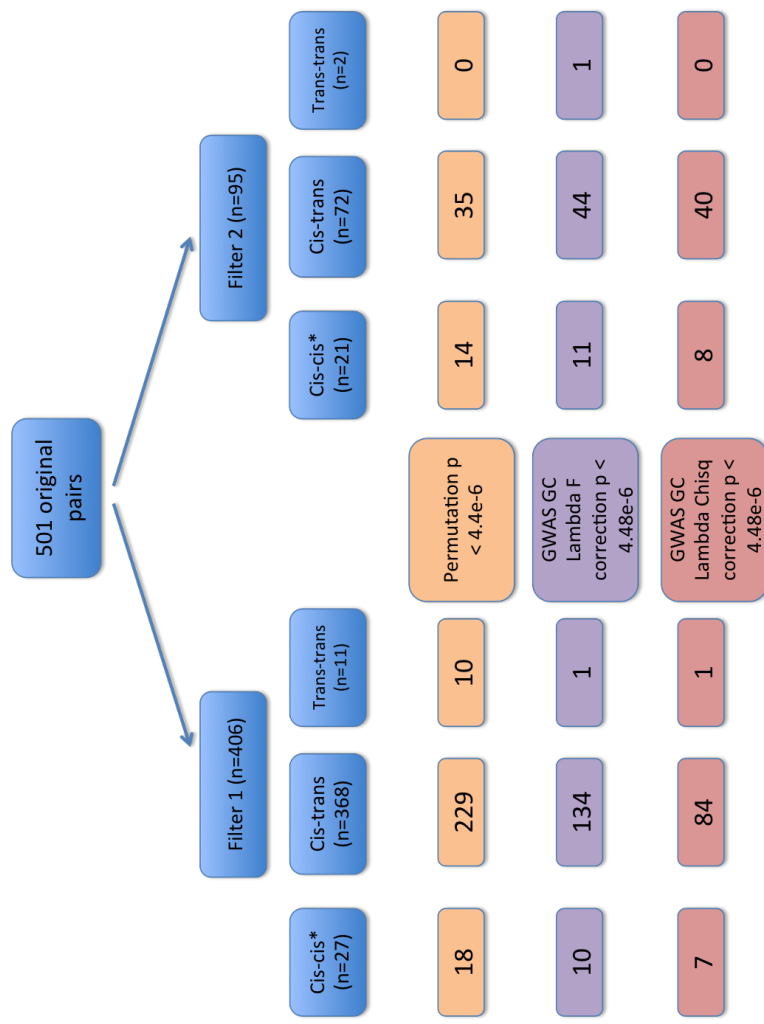
Figure | This figure denotes the numbers of Probe by SNP pairs that pass verious thresholds and split on original filtering and relative positional information. All pairs that pass both the permutation empirical p value threshold of $4.48e-6$ and the λ_F corrected threshold of $4.48e-6$ are summerized in the table *datasig.csv* and *datasig_lite.csv*. The total number of paris in this table is 155

15

probename	snp1	snp2	gene	chr1	chr2	lambda	P_emp	P_emp_ferh	P_emp_egcut
ILMN_1652333	rs898095	rs9892064	FN3KRP	17	17	1.34	6.77	7.21	7.11
ILMN_1710752	rs2123758	rs3889129	NAPRT1	8	8	2.14	7.36	7.42	6.99
ILMN_1731043	rs7776572	rs11770192	TRA2A	7	7	1.36	6.84	4.19	3.67
ILMN_1761797	rs9979356	rs3761385	CSTB	21	21	1.05	7.51	6.88	7.24
ILMN_1769782	rs1891432	rs10900520	LAX1	1	1	1.22	6.64	5.88	6.47
ILMN_1786426	rs807491	rs7254601	TMEM149	19	19	2.69	7.4	6.73	6.57
ILMN_1795836	rs9978658	rs11701361	C21ORF57	21	21	1.43	5.65	6.84	5.12
ILMN_1799381	rs2634462	rs6486334	SNORD14A	11	11	2.45	7.22	6.63	6.82
ILMN_2101920	rs6894268	rs4700810	HNRPH1	5	5	1.47	7.14	6.59	7.43
ILMN_2121437	rs7563453	rs4973397	NCL	2	2	1.2	5.88	5.87	5.96
ILMN_2242463	rs7930237	rs556895	CTSC	11	11	1.12	5.59	7.21	6.34
ILMN_2313158	rs16864367	rs13079208	MBNL1	3	3	2.28	7.22	6.84	6.35
ILMN_2358626	rs2395095	rs10824092	ADK	10	10	1.15	7.74	5.68	5.77
ILMN_2398939	rs8092433	rs4890876	MBP	18	18	1.19	5.9	6.59	6.11
ILMN_2410783	rs11150847	rs12602462	GAA	17	17	1.38	6.91	6.67	6.86

Summary information for the pairs of SNPs that had permutation empirical $p < 4.48^{-6}$. P_emp ferh and egcut are the permutation empirical p-values corresponding to the f-statistics from the original replication

Suggestions following discussions

1. 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)

2. New permutation

Explore a permutation analysis where $a + d$ for SNP1 and SNP2 are held constant. The suggestion was to permute the "residual" 4 df terms:

The model is:

$$y_i = a1 * xa1_i * xd1_i + a2 * xa2_j + [aa * xa1_i * x2a_i + ... + dd * xd1_i * xd2_i] + \epsilon$$

with xa and xd indicator variables (dummies). I think that the suggestion was to permute the entire term in square brackets (say, R_i) across individuals, holding y and the first 4 terms (dummies for a and d) constant. An alternative is to permute each of the 4 residual terms across y and $a + d$ dummies.

Response to specific email questions

Q1. How many pairs from the permutation analysis passed the Bonferoni threshold?

A1. Bonferoni threshold is $= 4.48^{e10-6}$. There are 305 / 501 pairs with empirical p-values from the permutation that are smaller

Q2. How many pairs that ended up in Table 1 of Hemani passed the Bonferoni threshold?

A2. 15

Q3. How many of the 95 pairs (filter 2) are above the B threshold and how many were in Table 1 of Hemani?

A3. 1

Q4. My guess (from your previous results) is that all cis-trans pairs are among the 30 that didn't pass the threshold - is that correct? Are there any cis-trans pairs among the 305 ones that survived the Bonferoni threshold with the new empirical p-value?

A4. Yes, that's correct. It's just cis-cis ones that pass the threshold. Although, of the 305 that do, 273 are cis-trans.