

Online methods

Contents

1	Discovery data	1
1.1	Data description	1
1.2	Normalisation	2
2	Exhaustive 2D-eQTL analysis	2
3	Replication	4
3.1	Data description	4
3.2	Meta Analysis	4
4	Additive vs epistatic variance comparison	5
5	Functional enrichment analysis	5
5.1	Tissue specific transcriptionally active regions	5
5.2	Chromosome interactions	6
5.3	SNP colocalisation with genomic features	7
5.4	Transcription factor enrichment	7
5.5	Defining previously identified SNP associations	8
6	References	8

1 Discovery data

1.1 Data description

The Brisbane Systems Genetics Study (BSGS) comprises 846 individuals of European descent from 274 independent families.¹ 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.² Standard quality control (QC) filters were applied and the remaining 528,509 autosomal SNPs were carried forward for further analysis.

Gene expression profiles were generated from peripheral 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 $\text{MAF} > 0.05$ within 1000 genomes data; did not map to a listed RefSeq gene; were not significantly expressed (based on a detection p -value < 0.05) in at least 90% of samples. After this stringent QC 7339 probes remained for 2D-eQTL mapping.

1.2 Normalisation

Gene expression profiles were normalised and adjusted for batch and polygenic effects. Profiles were first adjusted for raw background expression in each sample. Expression levels were then adjusted using quantile and log2 transformation to standardise distributions between samples. Batch and polygenic effects were adjusted using the linear model

$$y = \mu + \beta_1 c + \beta_2 p + \beta_3 s + \beta_4 a + g + e \quad (1)$$

where μ is the population mean expression levels, c , p , s and a are vectors of chip, chip position, sex and generation respectively, fitted as fixed effects; and g is a random additive polygenic effect with a variance covariance matrix

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

The parameters σ_a^2 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 1 is assumed to follow a multivariate normal distribution with a mean of zero. Residuals were normalised by rank transformation and used as the adjusted phenotype for the pairwise epistasis scan to remove any skewness and avoid results being driven by outliers. The GenABEL package for R was used to perform the normalisation.³

2 Exhaustive 2D-eQTL analysis

We used epiGPU⁴ 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 7339 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:

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

$$H_1 : \sum_{i=1}^3 \sum_{j=1}^3 (\bar{x}_{ij} - \mu)^2 > 0; \quad (4)$$

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 SNP 2. This type of test does not parameterize 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 when searching for a wide range of epistatic patterns, although will also include any marginal effects of SNPs which must be dealt with post-hoc.⁵

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 traits we established an experiment wide threshold of $T_e = 0.05/(N_* \times 7339) = 2.91 \times 10^{-16}$. This is likely to be conservative as it assumes independence between probes.

Following filtering on this threshold only SNP pairs with at least 5 data points in all 9 genotype classes were kept. We then calculated the LD between interacting SNPs 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. We then performed a second filtering screen that was identical to the first but an additional step was included where any SNPs that had previously been shown to have a significant additive or dominant effect ($p < 1.29 \times 10^{-11}$) was removed, creating a second set of 4751 filtered SNP pairs.

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 432 and 117 SNP pairs with significant interaction terms from filters 1 and 2, respectively.

All variance decompositions were calculated using the NOIA method.⁶

3 Replication

3.1 Data description

We have attempted replication of the 501 significant interactions from the discovery set using three independent cohorts; Fehrmann, EGCUT, and CHDWB. It was required that LD $r^2 < 0.1$ and $D'^2 < 0.1$ between interacting SNPs, and all nine genotype classes had at least 5 individuals present in order to proceed with statistical testing for replication in both datasets. We also excluded any putative SNPs that had discordant allele frequencies in any of the datasets. Details of the cohorts are as follows.

Fehrmann: $n = 1240$ The Fehrmann dataset⁷ consists of peripheral blood samples of 1240 unrelated individuals from the United Kingdom and the Netherlands. Some of these individuals are patients, while others are healthy controls. Individuals were genotyped using the Illumina 610 Quad platform. RNA levels were quantified using the HT12v3 platform.

EGCUT: $n = 891$ The Estonian Genome Center of the University of Tartu (EGCUT) study⁸ consists of peripheral blood samples of 891 unrelated individuals from Estonia. They were genotyped using the Illumina 610 Quad platform. RNA levels were quantified using the HT12v3 platform.

CDHWB: $n = 139$ The Center for Health Discovery and Well Being (CDHWB) Study⁹ is a population based cohort consisting of 139 individuals of European descent collected in Atlanta USA. Gene expression profiles were generated with Illumina HT-12 V3.0 arrays from peripheral blood collected from Tempus tubes that preserve RNA. Whole genome genotypes were measured using Illumina OmniQuad arrays. Due to the small sample size, most SNP pairs did not pass filtering in this dataset (27 SNP pairs remained) and so we have excluded it from the rest of the analysis.

3.2 Meta Analysis

The 4 *d.f.* interaction p -values for each independent replication dataset were calculated using the same statistical test as was performed in the discovery dataset. We then took the interaction p -values from EGCUT and Fehrmann and calculated a joint p -value using Fisher's method of combining p -values for a meta analysis as $-2 \ln p_1 - 2 \ln p_2 \sim \chi^2_{4d.f.}$. As in the discovery analysis, all gene expression levels were normalised using rank transformation to avoid skew or outliers in the distribution.

4 Additive vs epistatic variance comparison

To compare the relative contribution to the phenotypic variance of gene expression levels between additive and epistatic effects we are constrained by the problem that non-additive variance components for a phenotype cannot be calculated directly. Here, we only have SNP pairs that exceed a threshold of $p < 2.91 \times 10^{-16} = T_e$. A strong conclusion cannot be made about the genome-wide variance contribution, but we can compare the variance explained by SNP effects at this threshold for additive scans and epistatic scans.

In Powell *et al* 2012¹ an expression quantitative trait locus (eQTL) study was performed searching for additive effects in the same BSGS dataset as was used for the discovery here. Using the threshold T_e for the additive eQTL study, 453 of the 7339 probes analysed here had at least one significant additive effect. Assuming that the phenotypic variance for each of the probes is normalised to 1, the total phenotypic variance of all 7339 explained by the significant additive effects was 1.73%.

Following the same procedure, at the threshold T_e there were 238 gene expression probes with at least one significant pairwise epistatic interaction out of the 7339 tested. In total the proportion of the phenotypic variance explained by the epistatic effects at these SNP pairs was 0.25%.

5 Functional enrichment analysis

5.1 Tissue specific transcriptionally active regions

We employed a recently published method (<http://www.broadinstitute.org/mpg/epigwas/>)¹⁰ that tests for cell-type-specific enrichment of active chromatin, measured through H3K4me3 chromatin marks¹¹ in regions surrounding the 731 SNPs that comprise the 501 discovery interactions. We subsequently refer to these 731 SNPs as the IndexSNPs. A recent paper shows that H3K4me3 chromatin marks are the most cell-type-specific marks in terms of co-localization with previously published GWAS loci.¹⁰ The idea behind the method is that variants related to a particular phenotype may affect cell-type-specific gene expression by changing regulatory elements in cell types relevant to that phenotype. Hence, overlap between associated variants and chromatin marks should occur preferentially in the relevant cell type(s). Our analysis tested for enrichment of these chromatin marks in 34 different cell types from 4 different broad tissue types.

For each IndexSNP, a locus region was defined, bounded on either side of the IndexSNP by the most distant SNP within 250 kb that was in LD ($r^2 > 0.8$). For each SNP in a locus region regulatory activity scores were calculated as the height of the nearest H3K4me3 mark divided by distance from the SNP to the H3K4me3 mark. The SNP with the highest score within each locus region in a given tissue was designated the BestSNP, which served as the score representing that locus.

Cell-type-specificity scores per locus were estimated by normalizing BestSNP scores so that the sum of scores for a given locus across all cell types equaled 1. Cell-type-specificity scores per tissue were defined by summing normalized Best-SNP scores across all loci within a given tissue. 10,000 sets of SNPs (matched to the IndexSNP regions having the same total number of H3K4me3 peaks) were sampled (from among background SNPs provided with the software) to estimate null distributions of cell-type-specificity scores per locus and per tissue. P-values for cell-type-specificity scores summed across all BestSNPs for each tissue (the observed per-tissue score) were estimated as the proportion of random SNP sets with a per-tissue score exceeding the observed per-tissue score. We identified loci with BestSNP cell-type-specificity scores falling at or above the 95th percentile of the corresponding null distribution as demonstrating greater than expected specificity within a given cell type.¹⁰

5.2 Chromosome interactions

It has been shown¹² that different regions on different chromosomes or within chromosomes spatially colocalise within the cell. We shall refer to the colocalisation of two chromosome regions as a chromosome interaction. A map of all known pairwise chromosome interactions for K562 blood cell lines was recently produced,¹³ and we hypothesised that part of the underlying biological mechanism behind some of the 501 epistatic interactions may arise from chromosome interactions. We found that 44 of the putative epistatic interactions were amongst SNPs that were within 5Mb of known chromosome interactions. This means that SNP A was no more than 2.5Mb from the focal point of the chromosome interaction on chromosome A, and SNP B was no more than 2.5Mb from the focal point on chromosome B.

We performed simulations to test how extreme the observation of 44 epistatic interactions overlapping with chromosome interactions is compared to chance. Chromosome interactions fall within functional genomic regions,^{12,13} and the SNPs in our epistatic interactions are enriched for functional genomic regions. Therefore, we designed the simulations to ensure that the null distribution was of chromosome interactions between SNPs enriched for functional genomic regions but with no known epistatic interactions. To do this we used the 731 SNPs that form the 501 putative epistatic interactions and randomly shuffled them to create new sets of 501 pairs, disallowing any SNP combinations that were in the original set. Therefore, each new random set was enriched for functional regions but had no genetic interactions. We scanned the map of chromosome interactions for overlaps with the new sets and then repeated the random shuffling process. We performed 10,000 such permutations to generate a null distribution of chromosome interaction overlaps.

We repeated this process, searching for overlaps within 1Mb, 250kb, and 10kb.

5.3 SNP colocalisation with genomic features

We tested for enrichment of genomic features for the 687 IndexSNPs that comprise the 434 epistatic interactions with data present in discovery and replication datasets. For each of the 687 IndexSNPs we calculated LD with all regional SNPs within a radius of 0.5Mb and kept all regional SNPs with LD $r^2 > 0.8$. We then cross-referenced the remaining regional SNPs with the annotated chromatin structure reference¹⁴) querying whether the regional SNPs fell in Predicted promoter region including TSS (TSS), Predicted promoter flanking region (PF), Predicted enhancer (E), Predicted weak enhancer or open chromatin cis regulatory element (WE), CTCF enriched element (CTCF), Predicted transcribed region (T), or Predicted Repressed or Low Activity region (R) positions. Therefore a particular IndexSNP might cover multiple genomic features.

We then performed the whole querying process for each of the 528,509 SNPs present in the SNP chip used in the scan, and used the results from this second analysis to establish a null distribution for the expected proportion of SNPs for each genomic feature. We calculated p -values for enrichment of each of the seven genomic features independently, and for *cis*- and *trans*-SNPs separately, using a binomial test. For each genomic feature we used the expected proportion of SNPs as the expected probability of “success” (p). Here, a success is defined as an IndexSNP residing in a region that includes the genomic feature. The observed number of successes for each IndexSNP (k) out of the total count of IndexSNPs (n) was then modelled as $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$.

5.4 Transcription factor enrichment

To test for enrichment of transcription factor binding sites (TFBS) we followed a procedure similar to that described in Section 5.3. For each of the 687 IndexSNPs we extracted regional SNPs as previously described. We then used the PWMEnrich package in Bioconductor (<http://www.bioconductor.org/packages/2.12/bioc/html/PWMEnrich.html>) to identify which TFBSs each of the regional SNPs for one IndexSNP falls in (within a radius of 250bp). Thus, the number of occurrences of a particular TFBS was counted for each IndexSNP. We used the “Threshold-free affinity” method for identifying TFBSs.¹⁵

We constructed a null distribution of expected TFBS occurrences based on the same null hypothesis as described in Section 5.3 - the probability of an IndexSNP covering a particular TFBS is identical to any of the 528,509 SNPs in the discovery SNP chip. To do this, we performed the same procedure for each SNP in the discovery SNP chip as was performed for each IndexSNP to obtain an expected probability of covering a particular TFBS. We then tested the IndexSNPs for enrichment of each TFBS independently, and for *cis*- and *trans*-SNPs separately. p -values were obtained using Z-scores, calculated by using a normal approximation to the sum of binomial random variables representing motif hits along the sequence.¹⁶

5.5 Defining previously identified SNP associations

The discovery dataset (BSGS) had previously been analysed for additive and dominant marginal effects for all gene expression levels.^{1,17} To define SNPs that had been previously detected to have effects for a particular gene expression level we used a significance threshold accounting for multiple testing across SNPs and expression probes, $T_m = 0.05/(528509 \times 7339) = 1.29 \times 10^{-11}$. From this, we found that only nine of the 501 discovery interactions had known main effects, 64 were between SNPs that had no known marginal effects, and 439 were between a SNP with a known marginal effect and a SNP with no known marginal effect.

6 References

References

- ¹ Powell, J. E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS one* **7**, e35430 (2012).
- ² Medland, S. E. *et al.* Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American journal of human genetics* **85**, 750–5 (2009).
- ³ Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics (Oxford, England)* **23**, 1294–6 (2007).
- ⁴ Hemani, G., Theocharidis, A., Wei, W. & Haley, C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics (Oxford, England)* **27**, 1462–5 (2011).
- ⁵ Hemani, G., Knott, S. & Haley, C. An Evolutionary Perspective on Epistasis and the Missing Heritability. *PLoS Genetics* **9**, e1003295 (2013).
- ⁶ Alvarez-Castro, J., Le Rouzic, A., Carlborg, O., Álvarez Castro, J. M. & Carlborg, O. How to perform meaningful estimates of genetic effects. *PLoS Genetics* **4**, e1000062 (2008).
- ⁷ Fehrmann, R. S. N. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS genetics* **7**, e1002197 (2011).
- ⁸ Metspalu, A. The Estonian Genome Project. *Drug Development Research* **62**, 97–101 (2004).
- ⁹ Preininger, M. *et al.* Blood-informative transcripts define nine common axes of peripheral blood gene expression. *PLoS genetics* **9**, e1003362 (2013).
- ¹⁰ Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45**, 124–30 (2013).

- ¹¹ Koch, C. M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research* **17**, 691–707 (2007).
- ¹² Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* **326**, 289–93 (2009).
- ¹³ Lan, X. *et al.* Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic acids research* **40**, 7690–704 (2012).
- ¹⁴ Hoffman, M., Buske, O., Wang, J. & Weng, Z. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012).
- ¹⁵ Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
- ¹⁶ Ho Sui, S. J. *et al.* oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic acids research* **33**, 3154–64 (2005).
- ¹⁷ Powell, J. E. *et al.* Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLoS Genetics* **9**, e1003502 (2013).