Detection and replication of epistasis influencing transcription in humans

Gibran Hemani, Joseph E Powell, Konstantin Shakhbazov, Greg Gibson, Grant W Montgomery, Peter M Visscher

Abstract

A long standing question in evolution and human genetics is the extent to which epistasis, the phenomenon whereby one polymorphism's effect on a trait depends on other polymorphisms present in the genome, influences complex traits[1] and contributes to their variation.[2,3] Though epistasis has been demonstrated in artificial gene manipulation studies in model organisms,[4,5] and some examples have been shown in other species,[6] few convincing examples exist for epistasis amongst natural polymorphisms in human traits.[7,8] Its absence from empirical findings may simply be due to its unimportance in the genetic control of complex traits,[2,3] but an alternative view is that it has previously been too technically difficult to detect due to statistical power and computational issues.[9] Here we show that, using advanced computation techniques[10] and a gene expression study design, many instances of epistasis are found. In a cohort of 846 individuals with data on 7339 gene expression levels in whole blood, we found 501 significant pairwise epistatic interactions between single nucleotide polymorphisms (SNPs) acting on the expression levels of 238 genes ($p < 2.91 \times 10^{-16}$). Thirty interactions replicated in two independent datasets[11,12] following Bonferroni correction for multiple testing. Of the SNPs that did not pass this replication threshold, there was significant enrichment for interaction effects ($p < 1.0 \times 10^{-16}$). There was evidence of functional enrichment for the interacting SNPs, for instance 44 of the genetic interactions are located within 2Mb of regions of known intra-cellular chromosome interactions[13] ($p = 1.8 \times 10^{-10}$). Epistatic networks of three SNPs or more influence the expression levels of 129 genes, whereby one *cis*-acting SNP is modulated by several *trans*-acting SNPs. For example MBNL1 is influenced by an additive effect at rs13069559 which itself is masked by *trans*-SNPs on 14 different chromosomes, with nearly identical genotype-phenotype (GP) maps for each *cis-trans* interaction. This study presents the first evidence for multiple instances of epistatic genetic effects emerging from natural genetic variation in humans.

1 Main text

In the genetic analysis of complex traits it is usual for SNP effects to be estimated using an additive model where they are assumed to contribute independently, and cumulatively to the mean of a trait. This framework has been successful in identifying thousands of associations,[14] but to date there is little empirical exploration of the role that epistasis plays in the architecture of complex traits in humans,[7,8] though its contribution to phenotypic variance is frequently the subject of debate.[1–3] Outside the prism of human association studies there is evidence for epistasis, not only at the molecular scale from artificially induced mutations[4] but also at the evolutionary scale in fitness adaptation[15] and speciation.[16]

Methods are now available to overcome the computational problems involved in searching for epistasis, but its detection still remains problematic due to reduced statistical power. For example increased dependence on linkage disequilibrium (LD) between causal SNPs and observed SNPs,[17,18] increased model complexity in fitting interaction terms,[19] and more extreme significance thresholds to account for increased multiple testing[9] all make it more difficult to detect epistasis in comparison to additive effects. When genetic effect sizes are small, as is expected in most complex traits of interest,[14] the power to detect epistasis diminishes rapidly. There are two simple ways to overcome this problem. One is by using extremely large sample sizes;[20] another is by analysing traits that are likely to have large effect sizes. Because our focus was to ascertain the extent to which epistasis exists amongst natural genetic variation we designed a study around the latter approach and searched for epistatic genetic effects that influence gene expression levels. Transcription levels can be measured for thousands of genes. These traits are largely heritable but on average less polygenic than high level phenotypes,[21] thus it is expected that many genetic effects will be relatively large, maximising the chance at detecting epistasis, should it exist.

In our discovery dataset (Brisbane Systems Genetics Study, BSGS[22]) of 846 individuals genotyped at 528509 SNPs, we exhaustively tested every pair of SNPs for genetic interactions against each of 7339 expression traits in whole blood. After stringent filtering and multiple testing correction (Methods) we had identified 501 putative genetic interactions influencing 238 gene expression levels. Of the 501 discovery interactions, 434 had available data and passed filtering (Methods) in two independent replication datasets, Fehrmann[12] and the Estonian Genomics Centre University of Tartu (EGCUT),[11] in which we saw convincing evidence for replication. We used the summary statistics from the replication datasets to perform a meta analysis to obtain an independent p-value for the putative interactions, and 30 were significant after applying a Bonferroni correction for multiple testing (Table 1). These significant interactions exhibited remarkable similarity in GP maps between all three datasets (Figure 2).

Additionally, there was extreme enrichment for interaction effects among the discovery SNPs that did not reach the stringent significance threshold for replication (Figure 3). We observed that 316 of the remaining 404 discovery SNPs had replication interaction p-values exceeding the one-tailed 2.5% confidence

interval under the null distribution of no effects ($p << 1.0 \times 10^{-16}$, Supplementary Figure S1). The congruence of the epistatic networks in discovery and replication datasets is shown in Figure 1, demonstrating that these complex genetic patterns are common even across independent populations. A further replication was attempted using the Centre for Health Discovery and Wellbeing (CHDWB) dataset,[23] but only 185 of the SNP pairs passed filtering because the sample size was small ($n = 139$), and due to insufficient power we found no evidence for replication. It should be noted that although it is a necessary step to establish the veracity of the signals from the discovery set, replication of epistasis is theoretically difficult because the dependence on LD between observed SNPs and causal variants is on average three orders of magnitude higher than it is for independent additive effects.[17,18] Therefore these results are very encouraging with regards to the detection and replication of epistasis.

Though seldom the focus of association studies, SNPs with known main effects are often tested for additive $\times$ additive genetic interactions,[9] but our analysis shows that this is unlikely to be the most effective strategy for its detection. The majority of our discovery interactions comprised of one SNP that had a previously known association and one SNP that had no previous association in the dataset[21] (439 out of 501). Only 9 interactions were between SNPs that both had marginal effects while 64 were between SNPs that had no known marginal effects. Additionally, we observed that the largest epistatic variance component for the 501 interactions was divided amongst additive $\times$ additive, additive $\times$ dominance and dominance $\times$ dominance terms proportional to what is expected by chance ($p = 0.22$ for divergence from expectation). This is not surprising because the patterns of epistasis used for statistical decomposition are not designed to resemble biological function.[24]

We observed a wide range of significant GP maps (Figure 2) but the most common pattern of epistasis that we detected involved a *trans*-SNP masking the effect of an additive *cis*-SNP. For example, MBNL1 (involved in RNA modification and regulation of splicing[25]) has a *cis* effect at rs13069559 which in turn is controlled by 13 *trans*-SNPs and one *cis*-SNP that each exhibit a masking pattern, such that when the *trans*-SNP is homozygous for the masking allele the decreasing allele of the *cis*-SNP no longer has an affect. We observed that nine of the 14 masking SNPs are located in intronic regions (proportion of SNP panel in introns = 0.05, $p = 3.11 \times 10^{-9}$). Each of these interactions have evidence for replication in at least one dataset and six are significant at the Bonferroni level (Supplementary Figure S2).

Of the discovery interactions, 47 were *cis*-*cis* acting (both SNPs were on the same chromosome as the expression gene), 441 were *cis*-*trans*-acting, and 13 were *trans*-*trans*-acting. In total the 501 interactions comprised 781 unique SNPs, which we analysed for functional enrichment. We tested the SNPs for cell-type specific overlap with transcriptionally active chromatin regions, tagged by histone-3-lysine-4,3-methylation (H3K4me3) chromatin marks, in 34 cell types.[26] There was significant enrichment for *cis*-acting SNPs in haematopoeitic cell types only ($p < 1 \times 10^{-4}$ for the three tissues with the strongest enrichment after adjusting for multiple testing, Supplementary Figure S4). However *trans*-

acting SNPs did not show any tissue specific enrichment ($p > 0.1$ for all tissues, Supplementary Figure S4). This difference between *cis* and *trans* SNPs suggests that there is a range of molecular mechanisms by which epistasis might arise and the *cis*-SNPs may provide tissue specificity in these interactions. There is also strong enrichment for SNPs to be localised in enhancer regions[27] (Supplementary Figure S7). This enrichment is consistent for both *cis* and *trans* SNPs ($p < 1 \times 10^{-6}$). In particular, there was substantial enrichment for the GATA2 binding motif within 1kb of all epistatic SNPs, a known regulator of transcription in haematopoietic cells[28] ($p = 1 \times 10^{-40}$).

We also demonstrate a putative novel mechanism by which biological function can lead to epistatic genetic variance. It has been shown that different chromosomal regions spatially colocalise in the cell through chromatin interactions.[13] We cross referenced our epistatic SNPs with a map of chromosome interacting regions ($n = 96139$) in K562 blood cell lines[29] and found that 44 epistatic interactions mapped to within 2Mb ($p < 1.8 \times 10^{-10}$), (Supplementary Figure S6). Relatedly, there was significant enrichment of *trans*-SNPs being located within 250bp to kid/KIF22 binding sites ($p = 8.4 \times 10^{-28}$) which are involved in intracellular chromosome transport.[30]

Though we present many instances of epistasis, quantifying its relative importance to complex traits in humans remains an open question. In this study we are able to identify 238 gene expression traits with at least one significant interaction given our experiment-wide threshold. How does this compare to the number of traits controlled by additive effects? The BSGS dataset has been previously analysed for additive effects at all expression traits,[22] and if we take all the additive eQTLs that were significant at the epistatic threshold of $p < 2.91 \times 10^{-16}$ we find that 453 gene expression levels out of the 7339 analysed had at least one significant expression quantitative trait locus (eQTL). Therefore it can be argued that the number of instances of detectable epistasis are substantial.

However in terms of their contribution to complex traits a more important metric might be the proportion of the variance that the epistatic loci explain.[2] Ideally one would approach this question from a whole genome level[31] but this is intractable for non-additive variance components. Yet some inference can be made from the ascertained effects in these analyses and it is evident that additive variance is overall a larger component than epistatic variance, as has been argued previously.[2,3] Taking the additive effects detected in Powell *et al* (2012) at the $p < 2.91 \times 10^{-16}$ threshold, we calculate that in total they explain 1.73% of the total phenotypic variance of all 7339 probes. By contrast, the epistatic variance from the interacting SNPs detected in this study explain 0.25% of all phenotypic variance, approximately seven times lower than the additive variance. This is broadly in line with theory,[2] but there are three caveats to this comparison. Firstly, the ratio of additive to epistatic variance may differ at different effect sizes. Secondly, the power of a 1 *d.f.* test exceeds that of an 8 *d.f.* test. And thirdly, the non-additive variance at causal variants is expected to be underestimated by observed SNPs in comparison to estimates for additive variance, due to differences in the rate of decay of the estimate

of the genetic variance of the causal SNPs as LD decreases with the observed SNPs.

Overall, we have demonstrated that it is possible to identify and replicate epistasis in complex traits amongst common human variants. The functional analysis of the significant epistatic loci suggests that there are a large number of possible mechanisms that can lead to non-additive genetic variation. Further research into such epistatic effects may provide a useful portal to understanding molecular mechanisms and complex trait variation with greater clarity. With computational techniques and data now widely available the search for epistasis in larger datasets for traits of broader interest is warranted.

1.1 Methods Summary

We searched for pairwise epistasis exhaustively in the BSGS discovery dataset,[22] which comprises 846 individuals who are genotyped at 528509 autosomal SNPs and who have gene expression levels measured in whole blood samples for 7339 probes representing 6158 RefSeq genes. Recent hardware and software[10] advances made it possible to perform the 1.03×10^{15} statistical tests to complete this analysis. We used permutation analysis[32] to calculate an experiment-wide significance threshold of 2.91×10^{-16} at the 5% family-wise error rate (FWER). SNP pairs were modelled for full genetic effects, including marginal additive and dominance at both SNPs plus four interaction terms. Though we could have used a less complex model to improve statistical efficiency, we deemed it important to be agnostic about the type of epistasis that might exist, and therefore chose not to over-parameterise the test.[18,19] Because there are many large marginal effects present in these data it was necessary to perform several filtering steps to exclude SNP pairs that were significant due to marginal effects alone. All SNP pairs with LD $r^2 > 0.1$ were removed, and were required to have at least five data points in all nine genotype classes. If multiple SNP pairs were present on the same chromosomes for a particular expression trait then only the sentinal SNP pair was retained. Finally, a nested test contrasting the full genetic model against the marginal additive and dominance model was performed for each remaining SNP pair (Methods), resulting in 501 significant interactions after Bonferroni correction for multiple testing of the filtered SNPs. The significant SNP pairs were carried forward for replication in two independent datasets that used the same expression assays for analysing transcription in whole blood, the Fehrmann dataset[12] ($n = 1240$) and the Estonian Genome Centre University of the University of Tartu (EGCUT) dataset[11] ($n = 891$). Of these, 434 passed filtering in both replication datasets. A meta analysis on the interaction p-values from each replication dataset was performed to provide an overall replication statistic for each putative interaction.

2 Tables

				$-\log_{10} p$	
	Gene (chr.)	SNP 1 (chr.)	SNP 2 (chr.)	Discovery	Replication
1	ADK (10)	rs2395095 (10)	rs10824092 (10)	6.69	39.82
2	ATP13A1 (19)	rs4284750 (19)	rs873870 (19)	5.30	14.23
3	C21ORF57 (21)	rs9978658 (21)	rs11701361 (21)	9.42	21.67
4	CSTB (21)	rs9979356 (21)	rs3761385 (21)	11.99	42.27
5	CTSC (11)	rs7930237 (11)	rs556895 (11)	7.16	33.53
6	FN3KRP (17)	rs898095 (17)	rs9892064 (17)	16.16	59.95
7	GAA (17)	rs11150847 (17)	rs12602462 (17)	13.91	32.60
8	HNRPH1 (5)	rs6894268 (5)	rs4700810 (5)	15.38	10.37
9	LAX1 (1)	rs1891432 (1)	rs10900520 (1)	19.16	29.24
10	MBNL1 (3)	rs16864367 (3)	rs13079208 (3)	13.49	41.56
11	MBNL1 (3)	rs7710738 (5)	rs13069559 (3)	7.92	9.28
12	MBNL1 (3)	rs2030926 (6)	rs13069559 (3)	7.10	5.53
13	MBNL1 (3)	rs2614467 (14)	rs13069559 (3)	5.74	5.30
14	MBNL1 (3)	rs218671 (17)	rs13069559 (3)	7.63	5.23
15	MBNL1 (3)	rs11981513 (7)	rs13069559 (3)	7.71	4.58
16	MBP (18)	rs8092433 (18)	rs4890876 (18)	5.40	28.73
17	NAPRT1 (8)	rs2123758 (8)	rs3889129 (8)	8.45	30.77
18	NCL (2)	rs7563453 (2)	rs4973397 (2)	7.31	12.70
19	PRMT2 (21)	rs2839372 (21)	rs11701058 (21)	4.81	4.06
20	RPL13 (16)	rs352935 (16)	rs2965817 (16)	4.98	17.24
21	SNORD14A (11)	rs2634462 (11)	rs6486334 (11)	7.31	23.22
22	TMEM149 (19)	rs807491 (19)	rs7254601 (19)	12.16	145.78
23	TMEM149 (19)	rs8106959 (19)	rs6926382 (6)	5.80	10.72
24	TMEM149 (19)	rs8106959 (19)	rs914940 (1)	6.22	9.20
25	TMEM149 (19)	rs8106959 (19)	rs2351458 (4)	7.30	8.00
26	TMEM149 (19)	rs8106959 (19)	rs6718480 (2)	8.55	7.36
27	TMEM149 (19)	rs8106959 (19)	rs1843357 (8)	6.21	6.00
28	TMEM149 (19)	rs8106959 (19)	rs9509428 (13)	9.44	4.47
29	TRA2A (7)	rs7776572 (7)	rs11770192 (7)	8.23	4.09
30	VASP (19)	rs1264226 (19)	rs2276470 (19)	5.09	4.95

Table 1: Epistatic interactions significant at the Bonferroni level in two replication sets. All p-values are for 4 *d.f.* interaction tests (exact in Discovery and meta analysis for Replication).

3 Figures

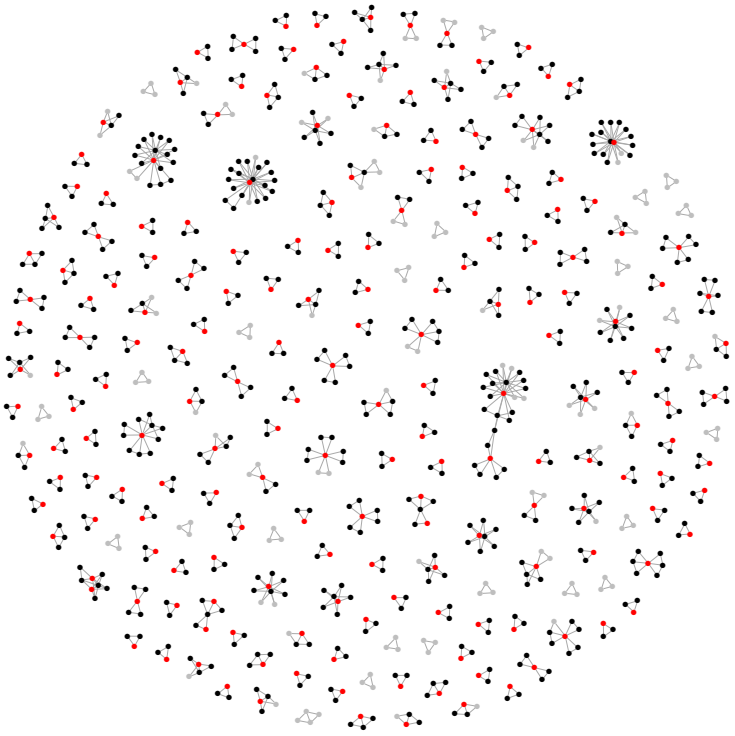

Figure 1: **Replication of epistatic networks in two independent cohorts** All 434 putative genetic interactions (edges) with data common to discovery and replication sets is shown. 316 Interactions that had p-values exceeding the 2.5% confidence interval following meta analysis of the replication data are shown with black nodes for SNPs and red nodes for gene expression traits. The interactions that fell within the confidence interval and showed no evidence of replication are greyed out. There is good congruence of interactions between independent populations.

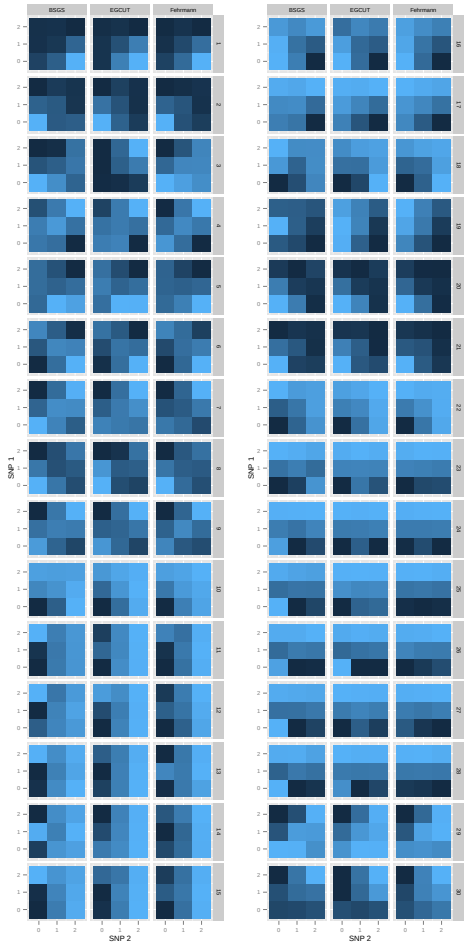

Figure 2: **Replication of genotype-phenotype (GP) maps in two independent populations** The GP maps for each epistatic interaction that is significant at the Bonferroni level in both replication datasets are shown. Each GP map consists of nine tiles where each tile represents the expression level for that two-locus genotype class. Phenotypes are for gene transcript levels (dark coloured tiles = low expression, light coloured tiles = high expression). Columns of GP maps are for each independent population. Rows of GP maps are for each of 30 significantly replicated interactions, corresponding to the rows in Table 1. There is a general trend of the GP maps replicating across all three datasets.

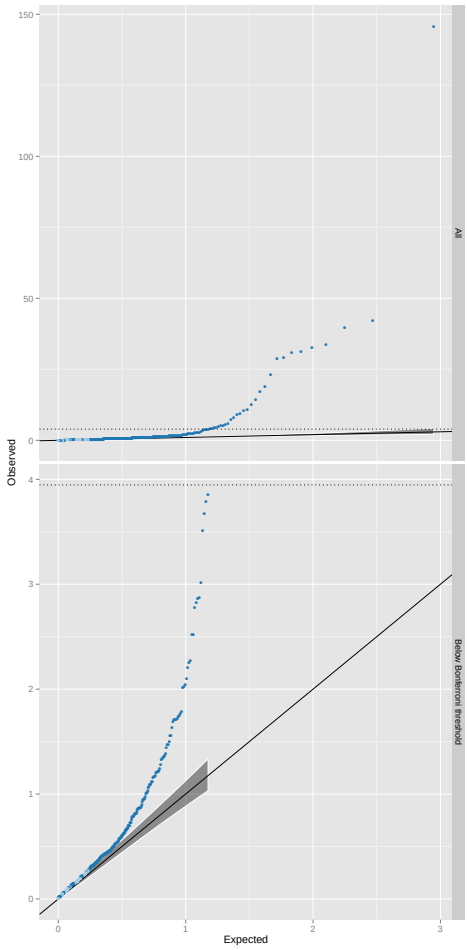

Figure 3: **Q-Q plots of interaction p-values from replication datasets** Top panel shows all 434 discovery SNPs that were tested for interactions. Observed p-values (y-axis, $-\log_{10}$ scale) are plotted against the expected p-values (x-axis, $-\log_{10}$ scale). The multiple testing correction threshold for significance is denoted by a dotted line. Bottom panel shows the same data as the top panel but excluding the 30 interactions that were significant at the Bonferroni level in the replication datasets. The shaded grey area represents the 5% confidence interval for the expected distribution of p-values. Dark blue points represent p-values that exceed the confidence interval, light blue are within the confidence interval.

4 Supplementary Figures

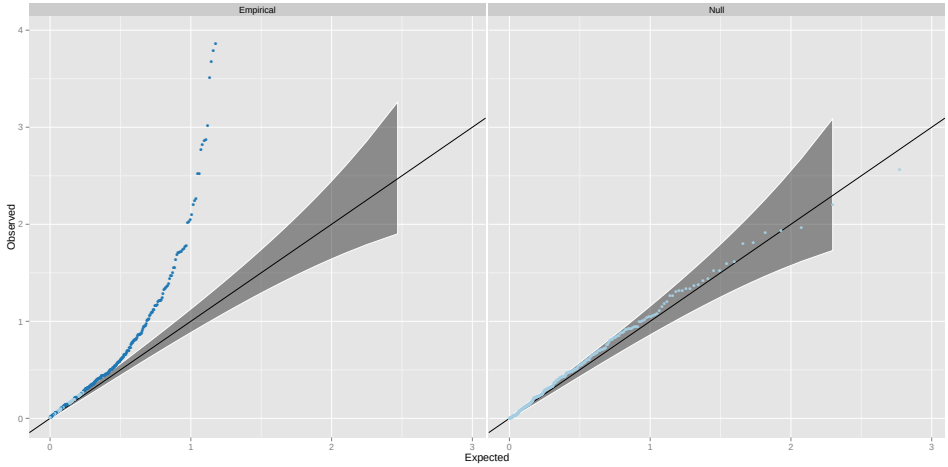

Figure S1: **Q-Q plots of interaction p-values from replication datasets, excluding the 30 points significant at the Bonferroni level** The right panel (Null) shows the interaction p-values from a meta analysis across two independent datasets on 434 randomly drawn SNP pairs. The left panel (Empirical) shows the interaction p-values from the 404 putative interactions that were not significant at the Bonferroni correction threshold. Dark blue points represent p-values that surpass the 5% FDR level, as in Figure 3.

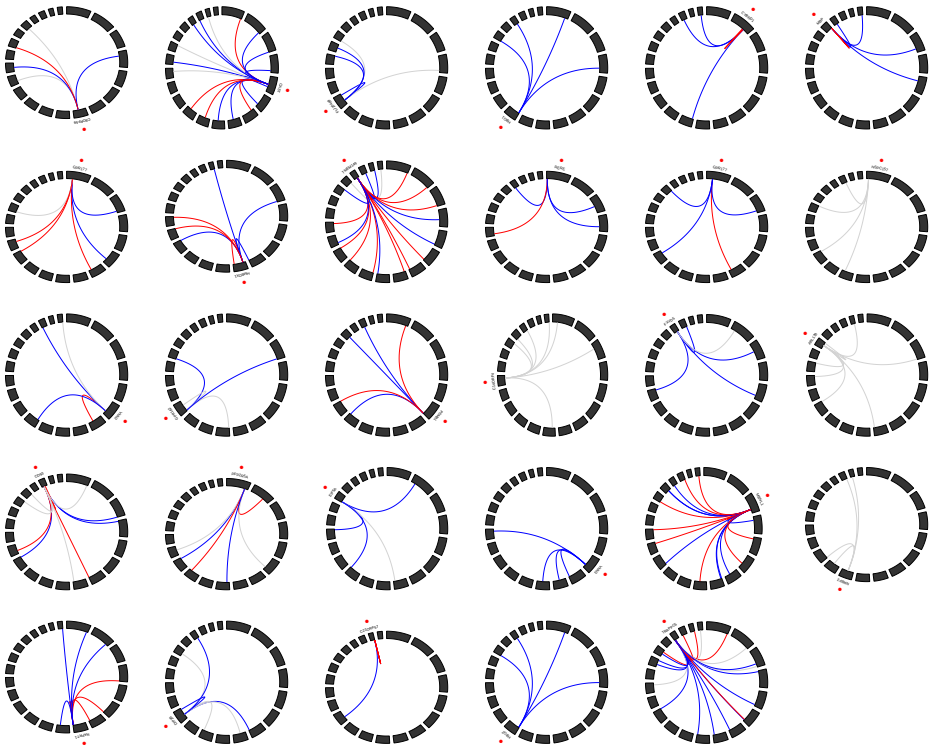

Figure S2: **Gene expression traits with four or more genetic interactions** Circle plots represent the genomic positions for SNPs (linking lines) and expression probes (red points). Chromosomes are represented by black blocks and ordered from 1 to 22 clockwise, starting from the top. Grey lines represent no evidence for replication, blue lines denote replication in at least one dataset, and red lines denote replication in two datasets. Most interactions are characterised as being *cis-trans* to the expression probe.

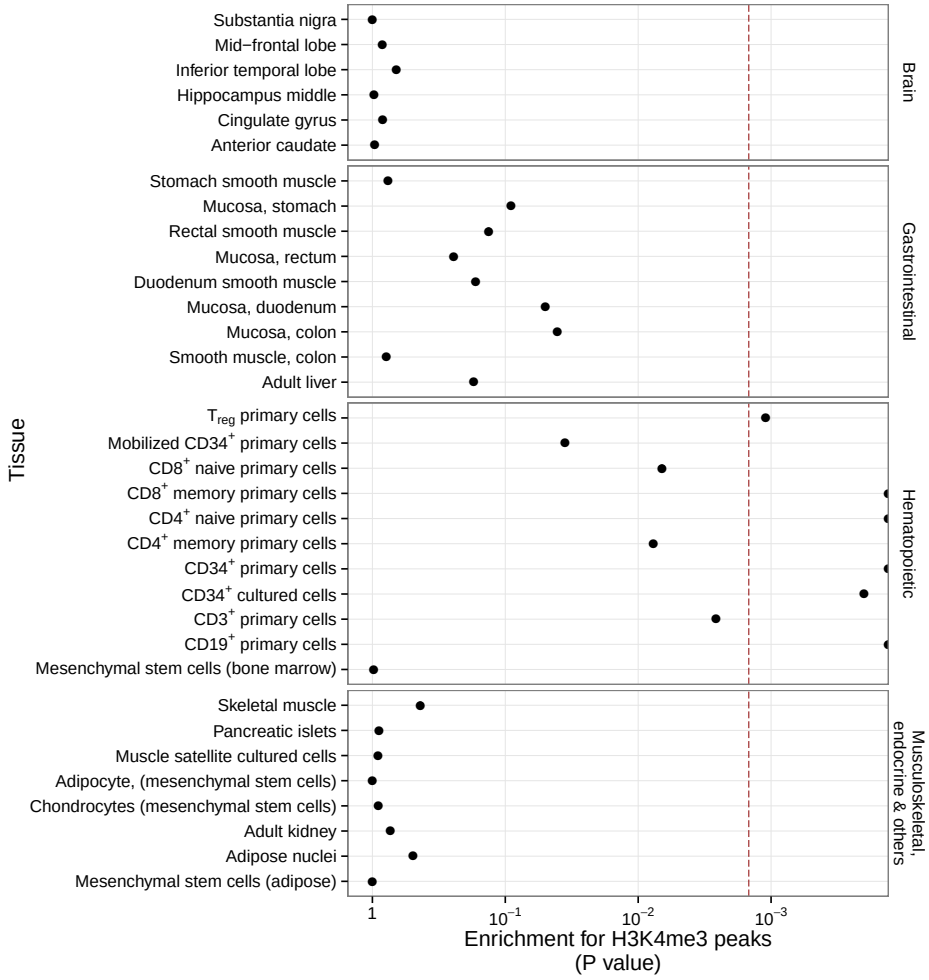

Figure S3: **Tissue specific enrichment of SNPs in transcriptionally active regions** The locations of transcriptional activity can be predicted by chromatin marks, assayed by H3K4me3. Here we show that there is

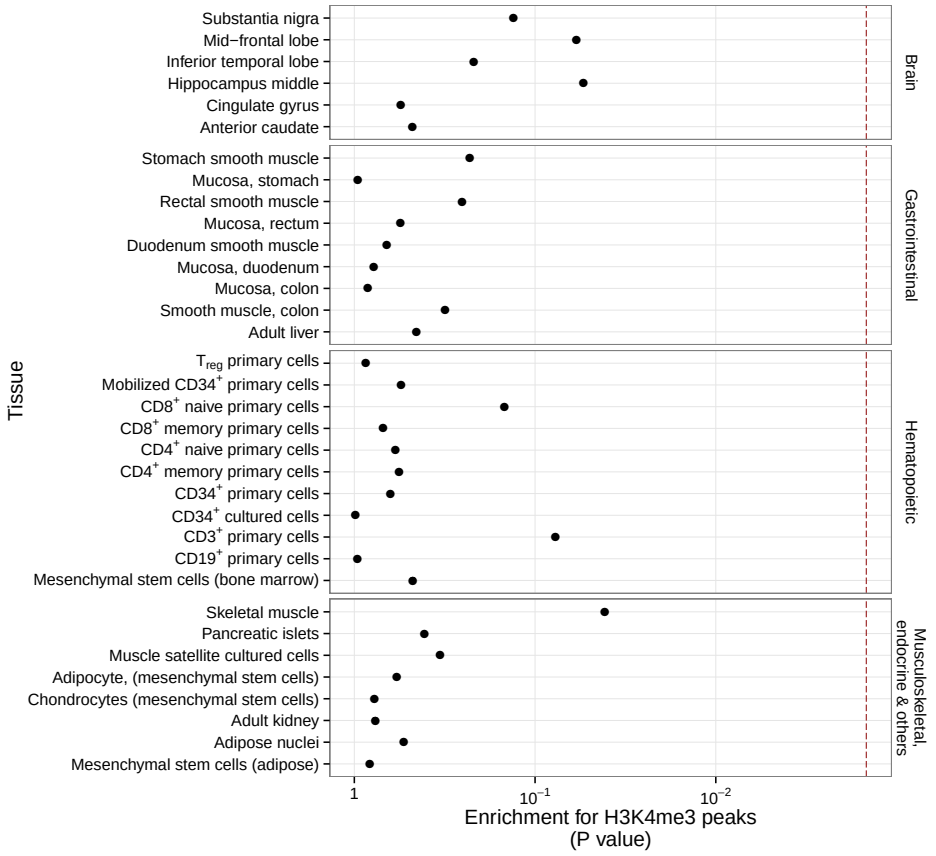

Figure S4: **Tissue specific enrichment of SNPs in transcriptionally active regions** The locations of transcriptional activity can be predicted by chromatin marks, assayed by H3K4me3. Here we show that there is

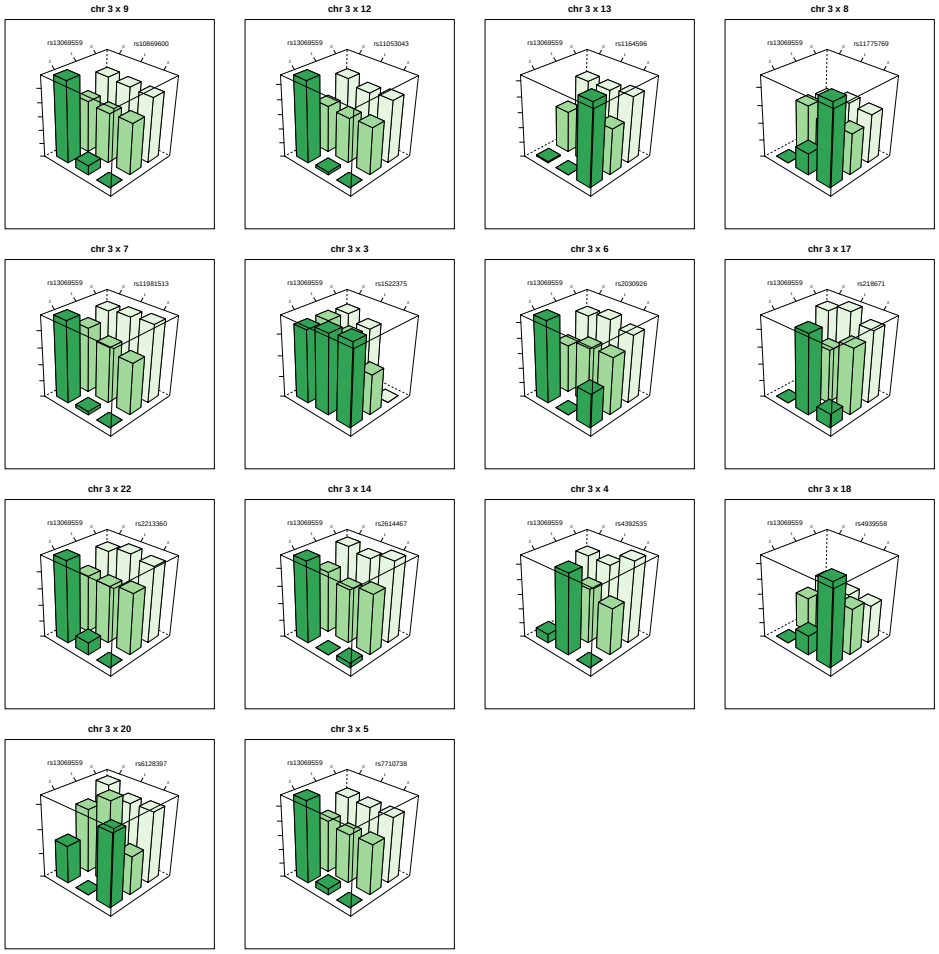

Figure S5: **Genotype-phenotype maps for 14 interactions controlling MBNL1** Each bar represents the mean phenotypic value for individuals in that genotype class. The rs13069559 SNP typically has a *cis*-additive decreasing effect on the expression of MBNL1, but in many of these interactions the *cis* effect is masked when the *trans* SNP is homozygous.

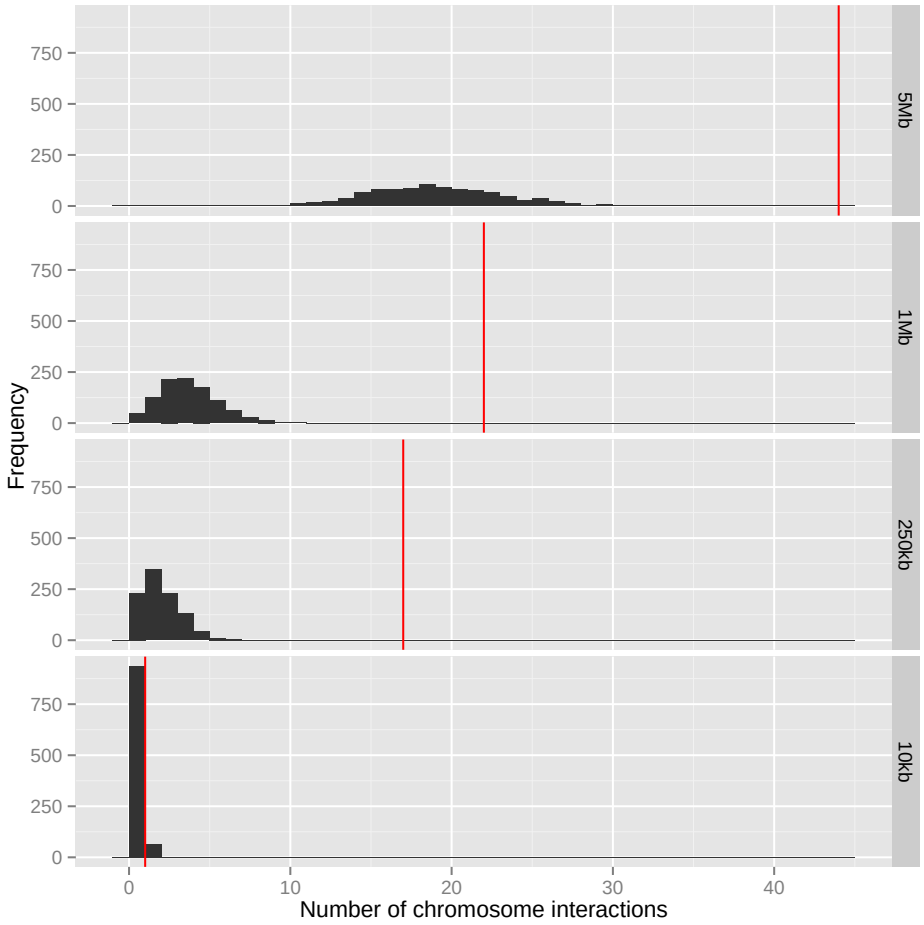

Figure S6: **Number of overlaps between chromosome interactions and epistatic interactions** Interacting chromosome regions may be a possible mechanism underlying epistatic interactions. The number of epistatic interactions within 20kb, 500kb, 2Mb and 10Mb of known chromosome interacting regions are shown by red vertical lines. The histograms represent the null distribution based on random sampling of 10000 datasets for each window size.

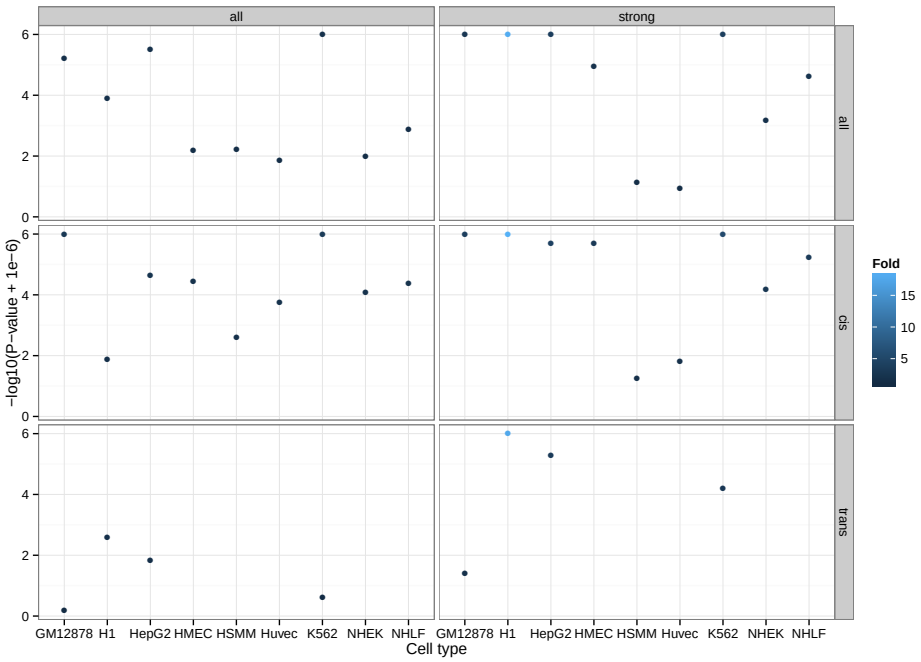

Figure S7: **There is enrichment for enhancer sequences for *cis* and *trans* SNPs**

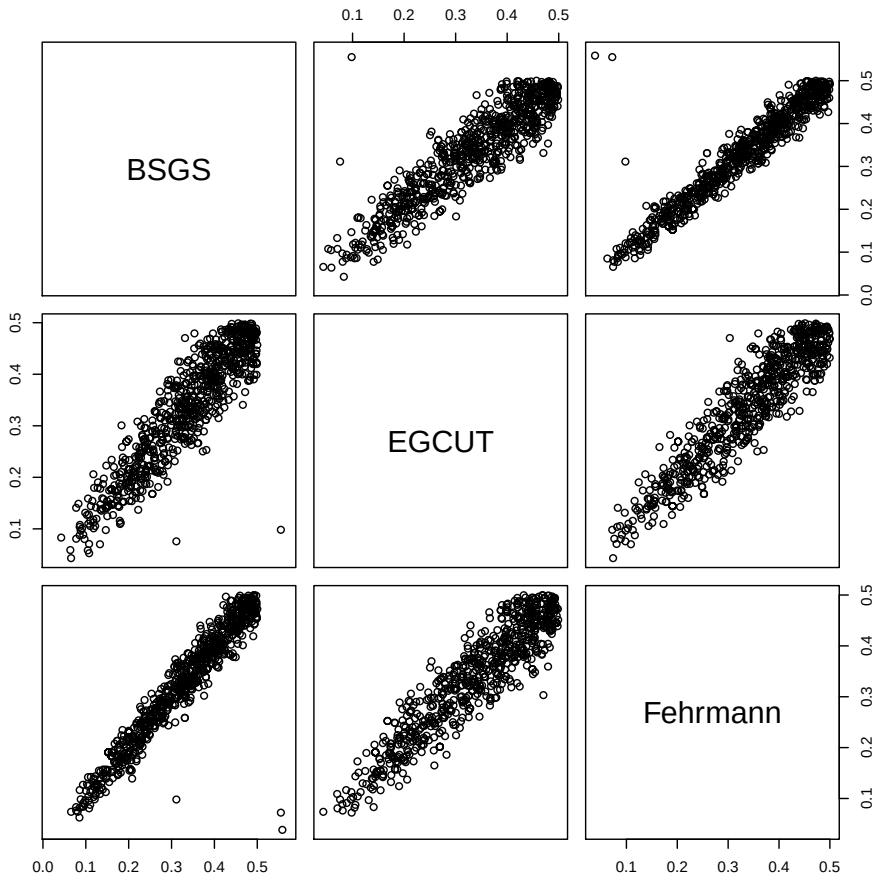

Figure S8: **Comparison of allele frequencies for 781 SNPs involved in genetic interactions across independent populations** Outliers were removed from the analysis as part of the filtering stage during replication.

5 References

References

[1] Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* **5**, 618–25 (2004).

[2] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics* **4** (2008).

[3] Crow, J. F. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365**, 1241–4 (2010).

[4] Costanzo, M. *et al.* The genetic landscape of a cell. *Science (New York, N.Y.)* **327**, 425–31 (2010).

[5] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. o. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* 1–6 (2013).

[6] Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P. & Andersson, L. Epistasis and the release of genetic variation during long-term selection. *Nature Genetics* **38**, 418–420 (2006).

[7] Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics* **42**, 985–90 (2010).

[8] Evans, D. M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics* **43** (2011).

[9] Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392–404 (2009).

[10] 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).

[11] Metspalu, A. The Estonian Genome Project. *Drug Development Research* **62**, 97–101 (2004).

[12] 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).

[13] 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).

[14] Visscher, P. M., Brown, M. a., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7–24 (2012).

[15] Weinreich, D. M., Delaney, N. F., Depristo, M. a. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science (New York, N.Y.)* **312**, 111–4 (2006).

[16] Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. a. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).

[17] Weir, B. S. Linkage disequilibrium and association mapping. *Annual review of genomics and human genetics* **9**, 129–42 (2008).

[18] Hemani, G., Knott, S. & Haley, C. An Evolutionary Perspective on Epistasis and the Missing Heritability. *PLoS Genetics* **9**, e1003295 (2013).

[19] Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417 (2005).

[20] Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–8 (2010).

[21] 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).

[22] Powell, J. E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PloS one* **7**, e35430 (2012).

[23] Preininger, M. *et al.* Blood-informative transcripts define nine common axes of peripheral blood gene expression. *PLoS genetics* **9**, e1003362 (2013).

[24] Cockerham, C. C. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882 (1954).

[25] Ho, T. H. *et al.* Muscleblind proteins regulate alternative splicing. *The EMBO journal* **23**, 3103–12 (2004).

[26] Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45**, 124–30 (2013).

[27] Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* **40**, D930–4 (2012).

[28] Tsai, F., Keller, G., Kuo, F. & Weiss, M. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221–226 (1994).

[29] 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).

[30] Miki, H., Setou, M., Kaneshiro, K. & Hirokawa, N. All kinesin superfamily protein, KIF, genes in mouse and human. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7004–11 (2001).

[31] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era–concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–66 (2008).

[32] Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–71 (1994).