

Epistatic control of gene expression in Humans

*** Report V1 ***

Joseph Powell
Gibran Hemani

April 24, 2013

1 INTRODUCTION

In humans expression quantitative trait loci (eQTL) have been extensively studied for their effects on transcript levels as well as their underlying effects on complex phenotypes such as disease susceptibility. To date most studies have analysed SNPs independently and estimated the genotypic effect on mean expression levels assuming an additive mode of inheritance. However, epistasis (gene by gene interactions) is thought to play an important role in the control of gene expression principally through the interaction of transcription factors and the intra-cellular effects of proteins on gene regulation.

The detection of epistasis is significantly hampered by power issues. For example, there is a greater dependency on linkage disequilibrium (LD) between causal variants and observed SNPs in order to capture non-additive variance, and this problem is exacerbated further with epistasis because there must be high LD at all interacting loci. Another obstacle comes from the dependence on a much larger sample size in order to detect similar effect sizes compared to additive effects, because of the increased number of parameters in statistical tests for interaction terms, and the increased multiple testing correction for testing combinations of SNPs. Indeed, most examples of epistasis in humans come from candidate gene studies where multiple testing does not have a detrimental effect.

Many approaches exist to detect epistasis, many of them are highly parametric, searching for a particular pattern of epistasis or only amongst SNPs with large marginal effects. However, it can be argued that the many patterns of epistasis will go undetected if the search is over-

parameterised, and it has been shown that using a test that is agnostic of the epistatic pattern in an exhaustive manner can detect signals that would otherwise go undetected. One method to overcome the problem of excessive multiple testing is to search for interactions in traits, such as gene expression, that are likely to have very large effects, thus improving the power of the study.

Recently we have shown that whilst most genetic variance for gene expression is likely to act in an additive manner, for many probes there exists significant non-additive genetic variance. The importance of epistasis for control of gene expression is largely unknown and has generally not been studied in human populations. It is likely that this is due in part to the considerable computation demands of running a 2D scan on 1000's of probes as well as a restrictive study-wide multiple testing burden. Due to modelling limitations, estimating the genetic variance due to non-additive genetic effects, and epistasis in particular, is often intractable. Thus, searching for epistatic effects directly may serve as a first step in quantifying the contribution of epistasis to complex traits.

2 METHODS

2.1 DISCOVERY DATA

The Brisbane Systems Genetics Study (BSGS) comprises 862 Individuals of European descent from 274 independent families [2]. 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. [1] Standard QC filters were applied and the remaining 528,509 SNPs were carried forward further analysis.

Gene expression profiles were generated from whole 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 MAF > 0.05 within 1000 genomes data; did not map to a listed ref-seq gene; were not significantly expressed (based on a detection p -value < 0.05) in 100% of samples. After this stringent QC 7000 [need to get exact number] probes remained for 2d-eqtl mapping.

2.2 NORMALISATION

Gene expression profiles were normalised and batch effects removed with the following procedures; Raw background expression were removed for each sample. Expression levels were then adjusted using Quantile and log2 transformation to standardise distribution between samples. To avoid inflation of the test statistic due to polygenic effects the following linear model was used to correct for batch and polygenic effects;

$$y = \mu + c + p + s + a + g + e \quad (2.1)$$

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

$$G_{ijk} = \begin{cases} \sigma_a^2 + \sigma_e^2 & j = k \\ 2\phi_{ijk}\sigma_a^2 & j \neq k \end{cases} \quad (2.2)$$

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 2.1 are assumed to follow a multivariate normal distribution with a mean of zero. These residuals are corrected for batch, sex, generation and polygenic effects. Residuals were standardised to Z-scores using an inverse-normal transformation and used as the adjusted phenotype for the pairwise epistasis scan.

2.3 EXHAUSTIVE 2D-EQTL ANALYSIS

We used epi-GPU 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 7000 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:

$$H_0 : \sum_{i=1}^3 \sum_{j=1}^3 (\bar{x}_{ij} - \mu)^2 = 0; \quad (2.3)$$

$$H_1 : \sum_{i=1}^3 \sum_{j=1}^3 (\bar{x}_{ij} - \mu)^2 > 0; \quad (2.4)$$

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 locus B. 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 (add refs), although will include any marginal effects of SNPs.

The complete exhaustive scan for 7,000 probes comprises of $\sim 9.8e^{14}$ F -tests. Due to the very large number of tests [plus nature of epistasis] we used a highly conservative series of filtering steps to identify study-wide significant epistasis associations to follow forward for replication analysis. Initially a significance threshold of $-\log_{10} 15.5$. This threshold is the study-wide significance threshold determined by the number of independent tests for a 2D scan for a single probe / the number of probes analysed. This is likely to be conservative as this threshold assumes independence between probes. Following filtering on this threshold only SNP pairs with all 9 genotype classes and minimum genotype class size > 5 individuals were kept. We then calculated the linkage disequilibrium between SNPs in a pair and removed any pairs with $r^2 > 0.1$. We also removed any pairs containing SNPs for which single marker additive or dominance effects had been identified in previous work (Powell et al. in Press). Sentinel SNP pairs we retained from epistasis eQTL 'peaks' comprising of multiple sets of pair-wise SNPs. At this stage $\sim 13,000$ SNP pairs remained and were tested for the contribution

of marginal and interaction terms by testing the 8d.f. 2.4 verses a 4 d.f. model parameterizing for interaction terms only:

$$H_0 : \sum_{i=1}^3 \sum_{j=1}^3 (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \mu)^2 = 0; \quad (2.5)$$

$$H_1 : \sum_{i=1}^3 \sum_{j=1}^3 (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \mu)^2 > 0; \quad (2.6)$$

where $\bar{x}_i(\bar{x}_j)$ is the marginal class mean for genotype $i(j)$ at SNP A (B). Significance of interaction terms was determined using a bonferroni threshold of $0.05/13000 = 3.8e^{-6}$

2.3.1 FINE MAPPING

To assess the impact of pair-wise epistasis variants not present on the Illumina 610Q beadchip and better define the association regions, we imputed genotypes in the region 100kb upstream and downstream of the sentinel SNPs using the full reference panel from the 1000 Genomes Project. The quality of imputation was assessed by means of the R^2 statistic. Used a highly conservative threshold, we removed any SNPs with $R^2 < 0.9$. Thus, for each epistasis pair two imputed SNP blocks were carried forward for fine mapping analysis. Each pairwise combination of SNPs in blocks 1 and 2 were tested in a full 8d.f. and reduced (interaction) 4d.f. model. As before any snp pairs not matching the criteria of $r^2 < 0.1$, number of genotype classes = 9, and minimum genotype class size > 5 were removed from the output.

This approach allows us to fine map to the causal epistatic loci. Given the nature epistasis genetic variance, any snp pairs with large estimate of non-additive genetic variance are very unlikely to be rare. Therefore, fine mapped loci provide good functional candidates. However, as fine mapping includes a large number of additional tests increases in the interaction (4d.f.) p -values could be due to chance [need a better way of describing this].

2.4 REPLICATION

We have attempted replication of the 2D scan significant results using three independent cohorts; Fehrmann, CHDWB and Rotterdam. Details of cohorts are as follows.

Fehrmann: $n=1,469$

The Fehrmann dataset consists of whole peripheral blood samples of 1,469 unrelated individuals from the United Kingdom and the Netherlands. Some of these individuals are patients, while others are healthy controls. Individuals were genotyped using Illumina HumanHap300, HumanHap370 or the 610 Quad platform. RNA levels were quantified using both the Illumina H8v2 platform ($N = 229$) and the HT12v3 platform ($N = 1,240$).

CDHWB: $n=189$

The Center for Health Discovery and Well Being (CDHWB) Study is a population based cohort consisting of 189 individuals collected in Atlanta USA. Gene expression profiles were

generated with Illumina HT-12 V3.0 arrays from whole blood collected from Tempus tubes that preserve RNA. Whole genome genotypes were measured using Illumina OmniQuad arrays.

Rotterdam: $n=762$

The Rotterdam Study is a large population based cohort collected to investigate the prevalence, incidence, and risk factors of chronic diseases among Caucasians aged 45 years and over. Participants were genotyped using the Illumina 610K quad arrays. Whole blood of 768 samples was collected (PAXgene) and RNA hybridized to the Illumina Human HT-12v4 chips.

2.5 FUNCTIONAL CHARACTERISATION

[From Kostya]

3 RESULTS

3.1 2D SCAN

We performed exhaustive two-dimensional searches on 7339 expression probes, identifying 432 significant pairwise interactions on 218 probes. We decomposed the genotype-phenotype map into parametric variance components and observed that there was no enrichment for any particular epistatic variance component amongst the significant SNPs (Table 3.1), demonstrating that a search that is agnostic of the pattern of epistasis is likely to detect things that would otherwise be missed. The majority of significant interactions were found to have a ‘cis-trans’ effect (Table 3.1; *i.e.* one SNP was on the same chromosome as the probe while the other was on a different chromosome). Of the 218 probes, 78 had more than one significant interaction (Figure 3.2)

	Interaction type	Freq
1	cis-cis	32
2	cis-trans	388
3	trans-trans	12

	Main epistatic effect	Freq
1	A x A	99
2	A x D	93
3	D x A	117
4	D x D	123

We attempted replication of the signals in an independent dataset [Greg Gibson - 139 individuals, 185 out of 432 interactions had both SNPs in common in the two datasets]. We observed no significant p -values after adjusting for multiple testing (lowest p -value= 0.0099 in replication dataset, threshold set at $0.05/185 = 0.0003$). We tested for enrichment of p -values but there was no evidence of departure from the null hypothesis of no effects (Figure 3.3).

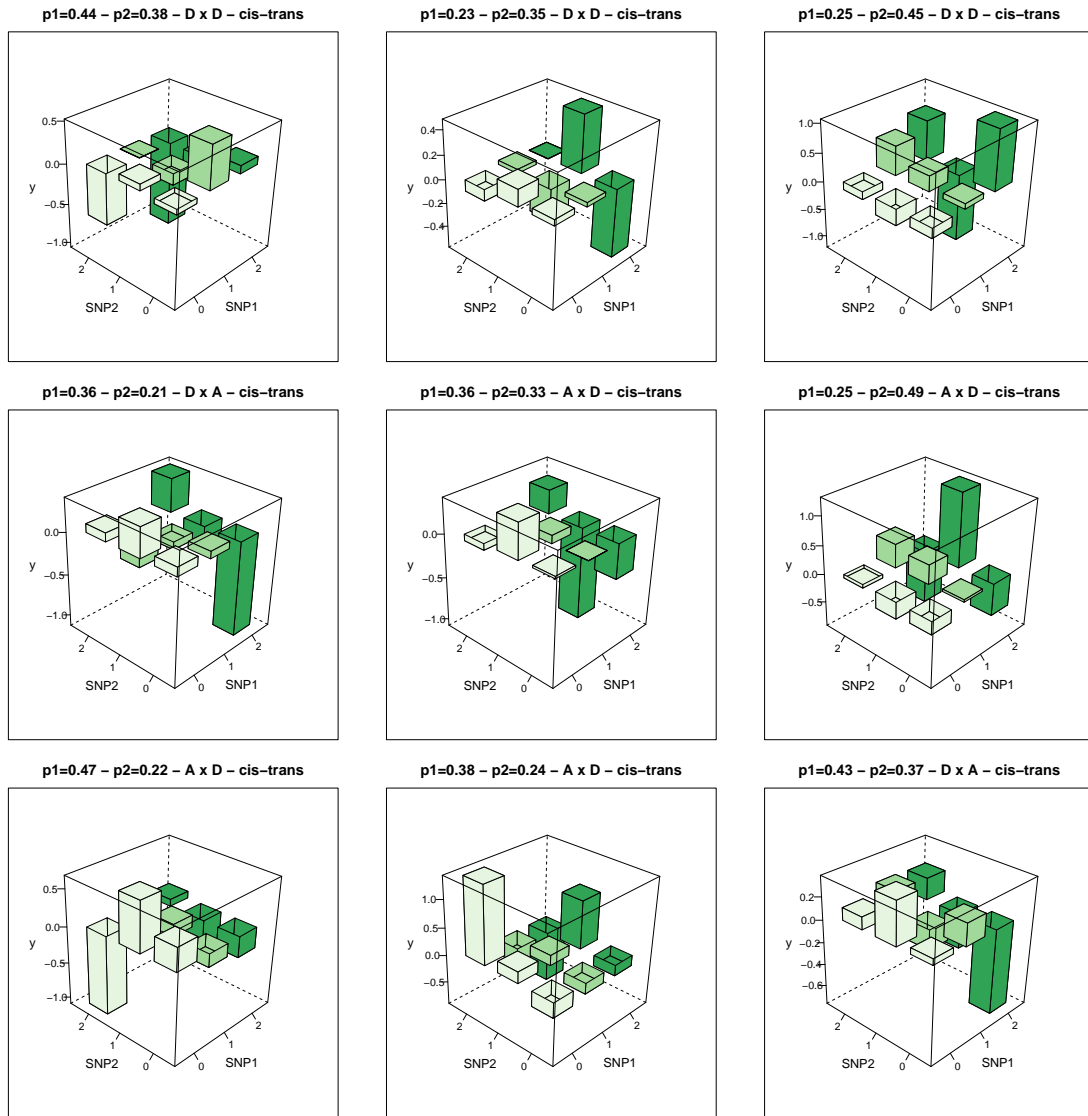


Figure 3.1: Genotype-phenotype maps for the nine most significant interactions.

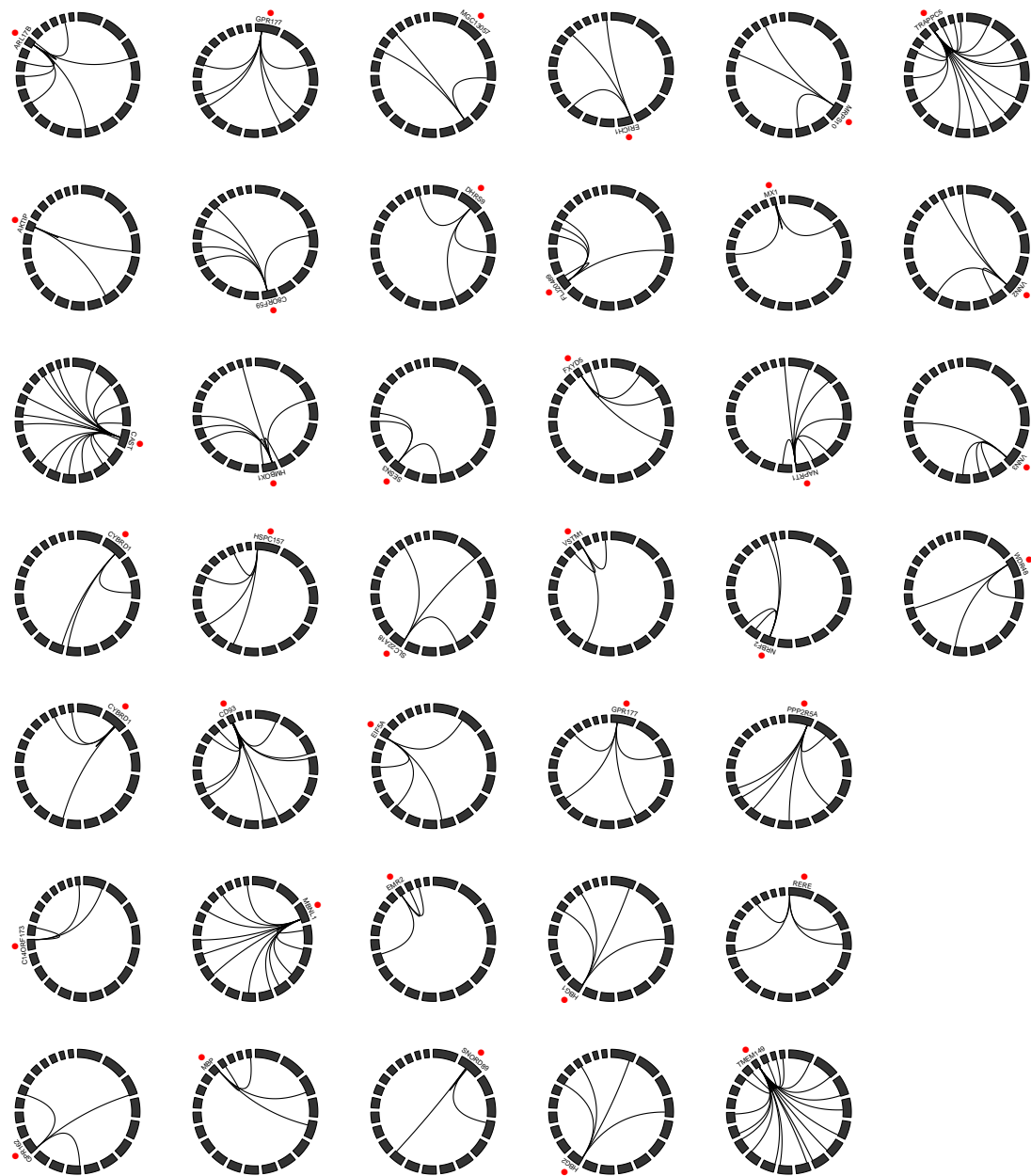


Figure 3.2: Genomic maps for 39 probes that have two or more significant interactions. The red dot denotes the chromosome where the probe resides and arcs denote interacting SNP pairs.

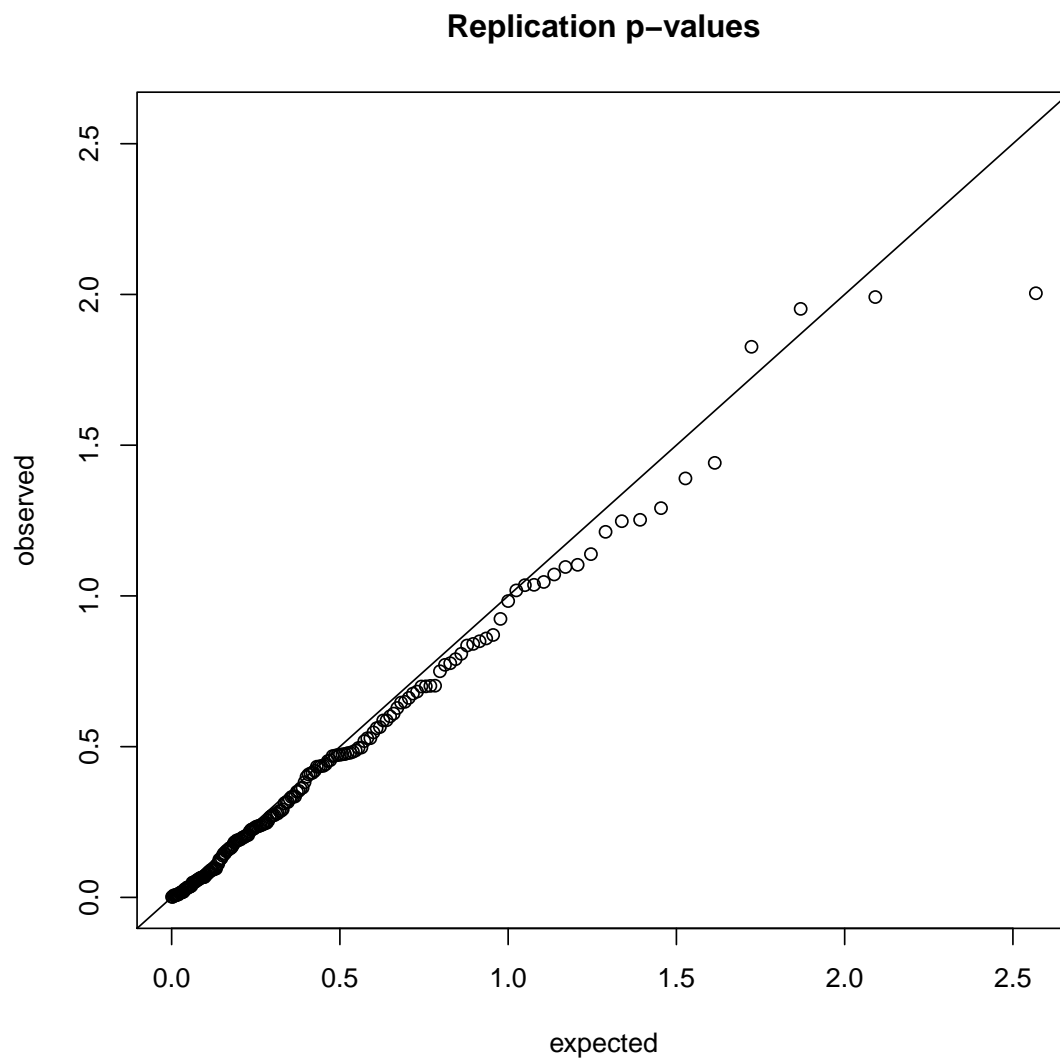


Figure 3.3: Q-Q plot for interaction p -values (4 degrees of freedom) for 185 interactions that had SNPs in common in the replication dataset.

3.2 ADDITIVE VS NON-ADDITIVE

This study has focused on identifying statistical interactions, however all significant interactions exhibit relatively large additive genetic effects (Figure 3.4).

3.3 FINE MAPPING

3.4 FUNCTIONAL CHARACTERISATION

REFERENCES

- [1] Sarah E Medland, Dale R Nyholt, Jodie N Painter, Brian P McEvoy, Allan F McRae, Gu Zhu, Scott D Gordon, Manuel a R Ferreira, Margaret J Wright, Anjali K Henders, Megan J Campbell, David L Duffy, Narelle K Hansell, Stuart Macgregor, Wendy S Slutske, Andrew C Heath, Grant W Montgomery, and Nicholas G Martin. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American journal of human genetics*, 85(5):750–5, November 2009.
- [2] J E Powell, A K Henders, A F McRae, A Caracella, S Smith, M J Wright, J B Whitfield, E T Dermitzakis, N G Martin, P M Visscher, and G W Montgomery. The brisbane systems genetics study: genetical genomics meets complex trait genetics. *PLoS ONE*, 7(4):e35430, 2012.

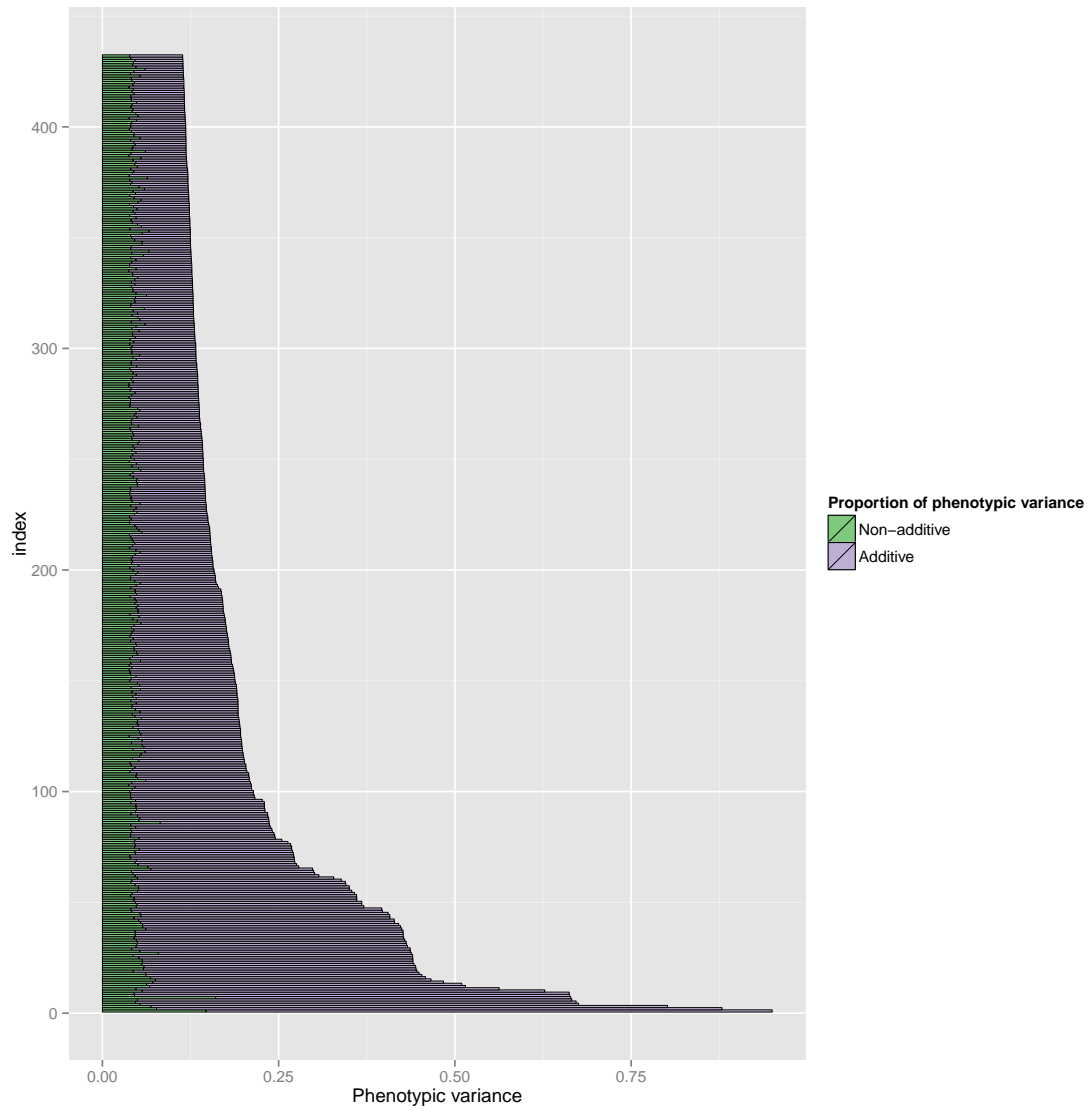


Figure 3.4: Proportion of genetic variance attributed to additive and non-additive effects, ranked in order of total genetic variance.