In model organisms, the widespread existence of trans eQTLs that regulate up to hundreds of genes,[1,2] has been well-established. However, in human genetic studies, relatively few examples have been detected. A major barrier to detecting these phenomena is arguably the traditional analysis approach, (call it Min-P), which considers only the smallest p-values and imposes a punishing multiple testing burden when applied in the context of a large number of tests for each variant. We can conceptualize a typical approach to trans eQTL mapping by thinking of testing each of t expression traits against each of m variants, resulting in mt tests, where t and m can be large. As an alternative to considering only the strongest signals, a large number of moderate or weak signals among the t tests for a given variant could potentially be aggregated to indicate the presence of a trans eQTL.

Variance component score test approach: One approach to address this problem is to use the idea of testing a variance component (VC) in a multi-trait LMM.

The corresponding test statistic would be $T = \sum_{j=1}^{t} Z_{ij}^2$ (could possibly cite[3,4]), in which the outcomes are the t (standardized) transcript levels, $Y_1, \ldots, Y_t$, and the predictor is the (standardized) genotype of variant i, G_i, where we allow random effects for genomewide SNPs and i.i.d. error. If we let α_{ij} denote the effect of G_i on Y_j and assume α_{ij} are i.i.d. $N(0, \sigma_i^2/t)$, then the idea would be to perform a score test of $H_0 : \sigma_i^2 = 0$ within the multivariate normal model with moments $E(Y_{n\times t}|X, G) = X\beta$, where $X\beta$ represents covariate and intercept fixed effects, and Var(vec$(Y)|X, G) = G_i G_i^T \sigma_i^2 + V_g \otimes K + V_e \otimes I$, where $K_{n\times n}$ is GRM and V_g and V_e represent $t \times t$ unknown cross-trait random effect covariance matrices (with a further constraint that the diagonal elements of Var(vec$(Y)|X, G)$ are 1). With individual-level data, the optimal score test under this model could be performed.

If instead we only had summary statistics, we could use the SPU(2) test[5] for multiple phenotypes, which is based on summary statistics and corresponds to testing whether the variance of i.i.d. random effects of a given variant G_i on each of the t expression traits, is zero. Suppose we have summary statistics Z_{i1},…,Z_{it}, where Z_{ij} is the score statistic (or similar score statistic) for testing Y_j on G_i in the simplified model $E(Y_j|G_i) = 1\beta_0 + G_i\gamma$ and Var$(Y_j|G_i) = I\sigma_e^2$. Then, we could set $T = \sum_{j=1}^{t} Z_{ij}^2$ and assess its significance under a simplified version of the full multi-trait model (with the constraints that there are no covariates and no additive genetic effects, but with V_e included in the model, now as a correlation matrix C_e), by comparing T to the null distribution that is as a linear combination of χ_1^2 statistics, where the linear coefficients are the eigenvalues of the covariance matrix of $(Z_{i1}, \ldots, Z_{it})$, where the latter can be estimated from the summary statistics under the simplified full trait model. The test based on T would be expected to be less efficient than the optimal score test based on individual-level data under the full trait model, but as usual, one must expect to pay a statistical price for analyzing summary statistics. A general feature of variance component score tests is that they have certain optimality properties in the case when the signals are extremely weak yet pervasive (not sparse). However, when the effects are very sparse and weak to moderate-sized, the variance component score test typically does not perform well (and this is supported by our preliminary results). For trans eQTL mapping, we have argued that sparseness is likely, so better statistical tools for that case are needed.

Methods for identifying variants that act epistatically on a phenotype The use of parallel computing has arguably made it feasible to perform tests of epistasis between every pair of variants for a given phenotype,[6,7] assuming access to suitable computing resources. However, even with those results in hand, there remain important statistical challenges in order to detect the signature of epistasis in the human genome.[8,9] There are some analogous methodology issues between the epistasis detection problem and the trans eQTL discovery problem discussed above. (Some of the comments we made about trans eQTL mapping above also apply to epistasis detection.) For example, in model organisms, the widespread existence of epistasis has been well-established.[10] However, in human genetic studies (which lack the advantages of experimental crosses), relatively few examples have been detected. In detection of epistasis, for each of m variants, one typically has $m - 1$ tests for epistasis, for a total of $m(m - 1)/2$ tests. In trans eQTL mapping, we framed the problem by considering an $m \times t$ matrix of test results, where interest lies in testing, for each row, the row-wise null hypothesis that all of the tests in that row are null. We argued that in the case of sparse and weak signals, there is a need for powerful approaches to combine the results in the row to perform a row-wide test. A similar argument has been made in the case of epistasis, where the idea of "marginal epistasis"[11] has been introducted. The idea is that rather than focusing on interacting pairs of variants, one could instead focus on one variant at a time and ask whether that variant has an overall epistatic signal for the trait. The idea is to aggregate many small epistasis signals for a given variant (from among the $m - 1$ tests it is involved in) to obtain overall evidence that the variant interacts. This non-traditional approach to analyzing epistasis aligns with recent methodological developments[12–15] in GWAS that attempt to learn about genetic architecture of complex traits by analyzing the many weak and moderate association signals throughout the genome.

References

[1] Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35: 57-64.

[2] Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7: 862–872.

[3] Zhou X, Stephens M (2014) Efficient algorithms for multivariate linear mixed models in genome-wide association studies. Nat Methods 11: 407-409.

[4] Dutta D, Scott L, Boehnke M, et al (2018) Multi-skat: General framework to test multiple phenotype associations of rare variants. Genet Epi Early Online.

[5] Kim J, Bai Y, Pan W (2015) An adaptive association test for multiple phenotypes with gwas summary statistics. Genet Epi 39: 651–663.

[6] Zhu S, Fang G (2018) Matrixepistasis: ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment. Bioinf 34: 2341-2348.

[7] Chatelain C, Durand G, Thuillier V, et al (2018) Performance of epistasis detection methods in semi-simulated gwas. BMC Bioinf 19.

[8] Wei WH, Hemani G, Haley C (2014) Detecting epistasis in human complex traits. Nat Rev Genet 15: 722–733.

[9] Ritchie MD, Van Steen K (2018) The search for gene-gene interactions in genome-wide association studies. Ann Transl Med 6: 157.

[10] Mackay TFC (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet 15: 22-33. PMC3918431.

[11] Crawford L, Zeng P, Mukherjee S, et al (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet 13(7): e1006869.

[12] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, et al. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47: 1228–1235. PMC4626285.

[13] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al (2015) An atlas of genetic correlations across human diseases and traits. Nat Genet 47: 1236-1241.

[14] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet 381: 13711379.

[15] Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. Cell 169: 1177-1186.