How to find epistasis, Bitches!

Gibran Hemani[1,2] and Joseph E Powell[1,2]

[1]University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [2]Queensland Brain Institute, University of Queensland, Brisbane, QLD, Australia

1 Abstract

2 Introduction

Epistasis is the process whereby the effect of one genetic loci on a phenotype is modified by the genotypes carried at other unlinked loci. As a topic in human genetics it has been widely discussed but rarely investigated in depth due to a series of computational and statistical challenges. Recently, a number of these challenges have been overcome and we have recently shown that through careful application of these methods (wide scale) epistasis can be detected for gene expression phenotypes in humans (ref). [I think we should try and say something about the processes needed to interoperate the output of a epistasis analysis]

2.1 Challenges

[Initially outline the challenges in the introduction. These can be address through the methods section]

1. Computational

[I think it would be worth re-discussing some of the computational problems of 2d scans. and highlighting the software + hardware solutions]

One of the first obstacles that make epistatic searches difcult is the computational demands of the statistical analyses. When searching for independent additive effects, as is done for the majority of GWA studies, each SNP is tested for association with the phenotype. However, in order to most powerfully identify epistatic effects, the search must be increased to multiple dimensions (ref). Here the scale of the computational demand increases by $(n^x)/2$ where n is the number of SNPs and x is the number of epistasis dimensions fitted. For example, testing for 2 loci interactions using SNP data from a 1 million SNP chip would require $1000000 * 999999/2 \approx 5e11$ individual tests.

2. Model choice

From a 2d 8df biallelic SNP model there are 4 epistatic variance components; additive x additive, additive x dominance, dominance x additive and dominance x dominance. Within the literature there has been considerable discussion as to the likely degree of epistasis variance components and whether statistical power could be increased by parameterizing models for only a subset. We have shown then estimates of parameters representing the 4 variance components are proportionally represented amongst the significant epistatic pairs. This implies a full model is preferential, particularly if we are agonistic about the mechanism by which epistasis might arise.

3. Statistical

A complete exhaustive scan of m phenotypes and n SNPs comprises of $((n * (n-1))^2/2) * m$ $8df$ F-tests. Given the high correlation structure inherent in genotype data as well as between multiple phenotypes, choices regarding multiple testing need to be carefully considered. [what can we say about tails of the distributions?] [Perhaps this is a good point to outline out a 2-step procedure where main effects are initially corrected for]
The 8df F-test includes parameters for main (additive) effects of the two SNPs. Therefore it is important to test that test statistic isn't driven my marginal effects - something that is quite likely for a trait such as gene expression. Thus one should really consider a nested test.

4. Interpretation / filtering

Things to consider and address during the analysis and output, particularly given the massive multiple testing and inability to easily evaluate the tails of the test-statistic distribution. Why we require at least 5 (or more) individuals in the smallest class size for the genotypes. The choice of number is partly based total sample size. Assuming Hardy-Weinberg equilibrium, the frequency of the smallest class size is $p_1^2 * p_2^2$ where p_1 and p_2 are the frequencies of the minor alleles for SNPs 1 and 2 respectively.

3 Methods

I think one approach for this paper (which is in keeping with other Nature Protocol manuscripts) is to go through the whole study design and show how we have done things - discussing choices in terms of the challenges mentioned above. For many steps we can show example (pseudo) code or give it in the supporting material

3.1 Stages of an epistasis analysis

3.1.1 Data

QC - probably want to use a high MAF cut-off than a standard GWAS in order to as anything with low MAF is going to increase the probability of $p_1^2 * p_2^2 * n$ being less than a required level.
Normalisation (distribution with respect to models used). Normal data. Is it worth discussing the potential problems caused by a skew in y and its effect on the tail of the test statistic distribution?

3.1.2 Computational / Statistics

Reiterate the distinction between CPU and GPU. I guess we can re-calculate some of the speed differences based on current GPU hardware and show speed

difference with relation to cpu based software. (Good place for a figure)

Outline the 2 stage models; 8df followed by (nested) 4df; or initial correction for main effects followed by a interaction term. This would also be the place to point out the importance of a) running a full 2d scan rather than initially search for marginal effects and b) including all epistasis variance component parameters in the interaction model.

3.1.3 Multiple testing

How to deal with multiple testing [what is the current state of your manuscript?]. Using GPUs it's not unrealistic that a researcher with access to a GPU cluster could run a permutation analysis for a single phenotype 2D scan. However, that should compare to your theoretically derived threshold. Therefore, the problem lies when we have multiple, correlated phenotypes such as expression.

3.1.4 Potential artifacts

Remove SNP pairs that are in LD greater than (threshold). Could lead in appearance of effects. Threshold determined based on....?

Population stratification and cryptic relatedness - look for inflation of LD between snp pairs.

4 Discussion