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, then address each in turn]

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 numer of snps and x is the number of epistatsis 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 epistatsis variance components and whether statistical power could be increased by parameterizing models for only a subset.

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?]

4. Interpretation / filtering
   The total output from an