**Supplementary Box S1: Rationale of frequentist methods for testing statistical interactions**

**Fisher’s definition**

In 1918, Fisher defined epistasis as a deviation from the addition of superimposed effects may occur between genes at different loci[1](#_ENREF_1). This is often referred to as ‘statistical epistasis’ to differ from the so called ‘biological epistasis’ first used in 1909 by Bateson[2](#_ENREF_2) describing a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect[3](#_ENREF_3). Fisher’s definition opened up a whole new avenue for using mathematical models to detect epistasis in complex traits at the population level, which may/may not correspond to biological consequences[3](#_ENREF_3). Mathematically, Fisher’s definition can be represented for two diallelic loci B (with alleles *b* and *B*) and C (with alleles *c* and *C*) by a linear model[3](#_ENREF_3), [4](#_ENREF_4):

(M1)



where *y* is a quantitative phenotype; *µ*, *aB*, *dB*, *aC*, *dC*, *iaa*, *iad*, *ida* and *idd* are the genetic parameters to be estimated, corresponding to the model mean, additive and dominance effects at the two loci and their epistatic interaction effects; *xi* and *zi* are dummy variables relating the genotype at locus *i* (B or C) with genetic effects, e.g. let *xB* = 1 and *zB* = -0.5 for a *b*/*b* genotype, *xB* = 0 and *zB* = 0.5 for *b*/*B*, and *xB* = -1 and *zB* = -0.5 for *B*/*B*. When ignoring interactions model M1 is reduced to:

(M2)



M1 is the saturated model fitting all nine possible parameters. One can use the 4 degree of freedom (df) test of interaction (i.e. *iaa* = *iad* = *ida* = *idd* = 0). Alternatively, interactions can be tested (4 df) by contrasting the maximum likelihoods of M1 (LS) and M2 (LR): i.e. if M1 fits data better interactions present or otherwise absent. When assuming alleles act additively without dominance within each locus, the interaction test will concern only *iaa* using 1 df.

Fisher’s definition can be applied to binary disease traits using logistic (instead of linear regression) models, in which case the quantitative phenotype *y* in M1 and M2 is replaced by where *p* is the probability of an individual being a case rather than a control in a population and LS and LR are maximum log-likelihoods of M1 and M2 respectively[3](#_ENREF_3), [5-7](#_ENREF_5).

**SNP genotype and allelic models**

In genome-wide association studies (GWAS), an exhaustive search for epistasis requires many billions of pairwise interaction tests but estimates of genetic parameters are necessary only for a small number of significant epistatic SNP pairs. A simple ‘genotype’ model fitting SNP genotypes directly (instead of *xi* and *zi* above) can focus on interaction tests and facilitate fast algorithm developments. Let *B* and *C* be the risk-modifying allele of locus B and C respectively, a genotype model can be represented in a 3 x 3 table (quoted from Supplementary information box S1 of an earlier review by Cordell[5](#_ENREF_5)):



where α represents the effect of the ‘baseline’ or reference genotype *b*/*b* *c*/*c*; β1 and β2 represent the effects replacing one or both *b* alleles at locus B with *B* allele; γ1 and γ2 represent the effects replacing one or both *c* alleles at locus C with *C* allele; *i11*, *i12*, *i21* and *i22* are interaction effects. The table above can be turned into the saturated logistic regression model (M3) below:

The 4-df tests above can be applied to test interactions in the genotype model. Assuming alleles act additively without dominance within each locus, allelic model (illustrated in a 2 x 2 table below) may be used and test the additive x additive interaction (*i*11):



**Contingency table based approximate tests**

The genotype/allelic models above require fitting the parameters against phenotypes (either a quantitative or binary trait) before performing the interaction tests but the fitting step is computational demanding. The (LS vs LR) test however can be approximated without the fitting step in order to greatly speedup genome-wise screening for epistasis. For example, for quantitative traits an F-ratio based test derived from two-way analysis of variance (ANOVA) appears to be a useful approximate test[8](#_ENREF_8) as illustrated below:



For a quantitative trait *y* with *n* samples, cells in a 3 x 3 contingency table represent nine joint genotypes each with *cntij* samples and a mean of *mij* of the trait (*i*/*j* is the row/column index, from 0 to 2). Based on the overall mean µ, row (*ri*) and column (*cj*) means, one can calculate variances (total: *SST*; between group: *SSB*; interaction: *SSI*; within group: *SSW*) and derive the F ratio test for interactions (*Fint*) using formula above (*dfI*: interaction df; *dfW*: within group df).

Similarly, for disease traits contingency table based approximate tests, e.g. Kirkwood Superposition Approximation (see Wan *et al.*[9](#_ENREF_9) for details), can be used to screen for SNP interactions, where two 3 x 3 contingency tables – one for cases and one for controls are considered (pij and qij are frequencies of each joint genotype in cases and controls respectively).



A good approximate test should be able to be computed fast without missing any potential interacting signals. It is critical to take an extra step of testing the resultant subset of potential interacting signals in the full regression models because: a) approximate tests often use relaxed conditions/assumptions and thus are subject to inflated false positives; b) in a genome-wide screening each approximate interaction test is conducted independently but potential correlations between tests could result in redundancy/dependency in the resultant interacting signals; c) interaction effects could be overestimated without considering the marginal effects of known loci confirmed in GWAS. The extra step is also applicable when the (LS vs LR) test itself is used because the (LS vs LR) test is not fully independent to the (LS vs LNull) test (i.e. testing if the saturated model is better than the Null model fitting no genetic parameters, LNull is likelihood of the Null)[10](#_ENREF_10) and false positives can occur in certain conditions[6](#_ENREF_6).

**Difference of inter-locus associations between cases and controls**

Based on the two 3 x 3 contingency tables above for diseases, the interaction parameters in model M3 can be estimated using log-odds under the assumption that Hardy-Weinberg Equilibrium (HWE) holds in the population and at least one of interacting SNPs under test has no marginal/main effects[6](#_ENREF_6):



This is an important property suggesting that testing interactions in logistic regression is equivalent to testing the differences of inter-locus associations between cases and controls and led to developments of a range of statistics modelling inter-locus association differences based on various definitions of either penetrance, or logit, or linkage disequilibrium, or haplotype (see detailed review and mathematic proofs in Hu *et al.*[6](#_ENREF_6) and Ueki & Cordell[7](#_ENREF_7)). Simulation results[6](#_ENREF_6), [7](#_ENREF_7) suggest that neither of these statistics is perfect for detecting epistasis in GWAS where the HWE assumption does not always hold and multiple SNPs with marginal effects do exist that are likely to be enriched in biological important pathways responsible for the disease studied. Nonetheless, these statistics at least qualify for good approximate tests. Again an extra of step of retesting the promising interacting SNPs in full logistic region models would be useful to remove potential false positives and construct the genetic structure underlying the disease by considering marginal effects of all known loci confirmed in GWAS and all epistatic interactions together.

1. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* **52**, 399-433 (1918).

2. Bateson, W. Mendel’s Principles of Heredity (Cambridge Univ. Press, Cambridge, 1909).

3. Cordell, H.J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**, 2463-8 (2002).

4. Cockerham, C.C. & Zeng, Z.B. Design III with marker loci. *Genetics* **143**, 1437-56 (1996).

5. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392-404 (2009).

6. Hu, J.K., Wang, X. & Wang, P. Testing gene-gene interactions in genome wide association studies. *Genet Epidemiol* **38**, 123-34 (2014).

7. Ueki, M. & Cordell, H.J. Improved statistics for genome-wide interaction analysis. *PLoS Genet* **8**, e1002625 (2012).

8. Gyenesei, A., Moody, J., Semple, C.A., Haley, C.S. & Wei, W.H. High-throughput analysis of epistasis in genome-wide association studies with BiForce. *Bioinformatics* **28**, 1957-64 (2012).

9. Wan, X. et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* **87**, 325-40 (2010).

10. Wei, W.H., Knott, S., Haley, C.S. & de Koning, D.J. Controlling false positives in the mapping of epistatic QTL. *Heredity (Edinb)* **104**, 401-9 (2010).