

TITLE: On Epistasis and Genome Scans for Complex Human Disease

AUTHOR: Susan R. Wilson

ADDRESS: Centre for Mathematics & its Applications
Mathematical Sciences Institute
The Australian National University
Canberra, ACT 0200, Australia

TEL: 61 2 6125 4460

FAX: 61 2 6125 5549

EMAIL: Sue.Wilson@anu.edu.au

RUNNING TITLE: On Epistasis and Genome Scans for Complex Human Disease

JOURNAL: Current Topics in Genetics

ABSTRACT

Epistasis, interaction among loci or between genes, is of growing interest and importance, especially as the quantity of genotype data being collected today is increasing at a dramatic rate. Most studies that attempt to identify the genetic basis of complex human disease ignore epistasis. As computer power has also been increasing dramatically, it is no longer necessary to be limited by a single-locus approach to the data. This paper reviews the underlying notion of epistasis and approaches that have been proposed for analysing multi-locus models for complex disease.

ABBREVIATIONS

QTL - quantitative trait locus/loci

SNP – single nucleotide polymorphism

KEYWORDS

Epistasis, gene-gene interactions, genome scan, complex human disease, QTL, linkage,

ACKNOWLEDGEMENT

The work was initiated while the author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2002. The visit was supported by the Institute and by a grant (No. 01/1/21/19/217) from BMRC-NSTB of Singapore.

INTRODUCTION

Epistasis and controversy have gone hand-in-hand since just after the dawn of research in Genetics a century ago. On one hand there is the biological reality of complex systems involving epistasis, and on the other hand the availability of data to statistically find epistatic terms involving interactions among loci. The well-known R.A. Fisher – Sewall Wright controversy on epistasis and its role in evolution continues [1], with Wright having emphasised the important role of epistasis in understanding the genetic basis of evolutionary change, while Fisher’s theory, that had assumed very large populations, concluded that the contribution of epistasis to evolution can be ignored.

There are two fundamental problems underpinning all debates. The first is definitional: What, exactly, is “epistasis”? The difficulties with its definition are overviewed in the following section. The second problem has been the type of data that has been collected until quite recently. The data issue is discussed in the sections on detecting epistasis and on recent, relevant research. The discussion looks towards the future.

EPISTASIS – DIFFICULTIES WITH ITS DEFINITION

Although epistasis is a basic concept in genetics whose use can be traced back to Bateson in 1907 [2], there has been confusion in the use of the term. This confusion has been discussed in detail elsewhere [3, 4]. In Bateson’s use of the term that is still widely used by biologists, epistasis is a masking effect by which one Mendelian locus alters the allelic effects at another locus. As is shown in [4], the classical ‘heterogeneity model’ falls within this definition. The inherent problem is lack of a precise (biological) definition.

The other use of the term that pervades quantitative genetics literature can be traced back to Fisher’s influential 1918 paper [5]. Here epistasis is defined as a deviation from additivity in the effect of alleles at different loci. Mathematically, this can be written

$$y = \mu + \sum_j (a_j x_j + d_j z_j) + \sum_{j < k} (i_{jk}^{aa} x_j x_k + i_{jk}^{ad} x_j z_k + i_{jk}^{da} z_j x_k + i_{jk}^{dd} z_j z_k)$$

where y is a quantitative phenotype (or penetrance if the trait is binary) and x_j and z_j are dummy variables related to the underlying genotype at locus j . The coefficients μ , a_j and d_j represent the mean effect, and the additive and dominance effects at locus j , and the i coefficients correspond to epistatic interaction effects. If there is no epistasis, all the i (interaction)

coefficients are zero and the model is said to be additive. Although this is a precise definition, a difficulty in practice is that models may be additive on one scale, but not additive, i.e. contain interaction effects (epistasis) on the scale of biological interest. Moreover, statistical interaction does not necessarily have biological meaning, i.e. it does not imply biological interaction or even, perhaps, any biological effect at all [3, 4].

DETECTING EPISTASIS IN COMPLEX (HUMAN) DISEASE

In analysis of real data for complex human disease there have been a few demonstrations of epistasis (gene-gene interactions), including type 1 diabetes [6], type 2 diabetes [7] and inflammatory bowel disease [8], and very recently prostate and breast cancer [9, 10].

Methods for the detection of epistasis vary according to whether one is using a linkage method or an association method. Many approaches for linkage analysis of multi-locus (initially two-locus) models have been proposed for the study of complex disease, and include likelihood-based linkage analysis [see, for example, 11], and non-parametric linkage [see, for example, 8, 9], permutation tests [12], allele-sharing-based linkage analysis [see, for example, 6, 13], marker-association-segregation method [14], weighted-pairwise correlation method [15], variance component analysis [7, 16] and recurrence risk of relatives [17]. Papers dealing with various two-locus models usually considered only a few (for example, 6 are typically used in linkage analysis [17]) of the 48 possible minimum set of nonredundant two-locus models (excluding zero-locus and single-locus models). Many of these additional 42 models may not be relevant to gene-gene interactions underlying complex disease, but even so, it is fairly straightforward to construct a biochemical system based on each of them [18].

There have been discussions of whether single-locus models suffice for detecting linkage even in the presence of epistasis; see, for example, [18]. It would seem to depend on the approach chosen, as some show significantly increased power to detect epistasis when two (or more) loci are considered simultaneously (see, for example, [11, 13]), while others apparently do not [16]. However, contrary to the results in [16], it has been shown that epistasis between two regions can be detected successfully by calculating the correlation between two linkage signals, each determined by a single-locus linkage analysis [7, 18].

Two of the major approaches for association analysis are for case-control data (see, for example, [10]) and analyses of affected individuals and their parents [see, for example, 19, 20]. Although logistic regression is widely used, there can be inherent problems arising from sparseness if a large number of loci are being simultaneously investigated, and so a nonparametric multifactor-dimensionality reduction method has been suggested as an alternative approach [21].

Results of association studies of common complex diseases do not replicate across different studies; see references in [22]. One reason is attributed to epistasis. Simpson's paradox, the reversal of the direction of an association by collapsing tables is well-known in applied statistics. This may well be what is happening when single-locus methods (i.e. for collapsed tabular data) are being used while the underlying genetic model involves two or more loci interacting epistatically.

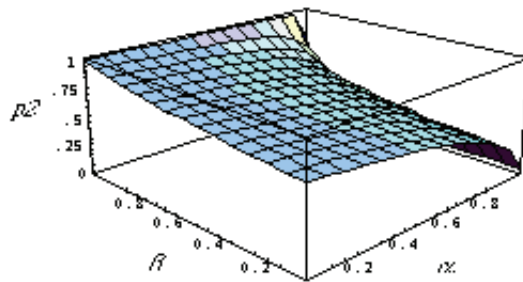


Fig. 1 For a given α and β , either preferred transmission of a or A will be inferred, dependent on whether p_2 falls above or below the value on the plane.

For example, consider the simplest case of two (unlinked) disease susceptibility loci, where each locus has just two alleles, and Hardy-Weinberg equilibrium holds. There are a total of 8 parameters (compared with only two for the single-locus two allele model). Assume the frequency of allele a at locus 1 is p_1 and of allele b at locus 2 is p_2 . Let the penetrances of those who are bb Bb BB be f αf αf , respectively for those who are aa , while they are βf f f for those who are Aa or AA , where f can be chosen to give the appropriate population disease frequency. It is straightforward to show for case-control data (by comparing the frequency of a is cases and controls) that allele a will appear to be associated with disease (cases) if $p_2 > ((1-\alpha)/(2-\alpha-\beta))^{1/2}$, and allele A if the inequality is reversed; see Fig. 1. Note that for all $\alpha=\beta$, the value on the plane given in Fig. 1 is $1/\sqrt{2}$. This result also has

been shown for analyses of affected individuals and their parents [20].

It can be shown using the approach outlined in [23] that, for the above simple example of two-locus penetrances, all the variation will be epistatic for variance component analyses if $p_2 = ((1-\alpha)/(2-\alpha-\beta))^{1/2}$, and that a single-locus approach will have no power at all, and little power for values near this equality (namely near the plane in Fig. 1). Further, it was shown in [22] that for purely epistatic models (with no additive or dominance variation at any of the susceptibility loci), association models analysing one locus at a time had no power, while linkage methods do have power due to the increased allele sharing between affected relatives.

Recently it was shown using simulated data that it is computationally feasible to look explicitly for statistical interactions between pairs of loci within a set of hundreds of thousands of genotyped markers, using logistic regression where the number of cases and of controls is each of order $O(10^3)$ [24]. Even with a conservative correction for multiple testing (Bonferroni), this approach was more powerful than traditional analyses for three biologically motivated parametric models of gene-gene interactions.

SOME RECENT, RELEVANT RESEARCH FOR QTLS

Epistatic quantitative trait loci (QTL-) mapping studies in model organisms have detected many novel interactions, with the resultant conclusion that epistasis makes a large contribution to the genetic regulation of complex traits [25]. Further, it is reasonable to assume that the importance and frequency of gene interactions in natural populations (including humans) are of the same magnitude as those found in model organisms. In [26] it is noted that progress has been made in identifying major QTLs but experimental constraints have limited our knowledge of small-effect QTLs, and these may actually be responsible for a large proportion of trait variation.

Using spotted microarrays, all transcript levels in a cross between two yeast strains were examined recently [27]. It was found that genetic interactions underlie the inheritance of about half of all these transcript levels, and that at least one member of an interacting locus pair had too small an individual effect to be identified on its own. Adjustment was made for multiple testing. However, the strategy used here would not have detected interacting locus pairs in

which both individual effects are small. Also, the interactions were investigated on the log scale for the expression measurements, and these may, or may not, correspond to interactions on the scale of biological interest, namely concentration.

DISCUSSION

Most genetic modelling strategies for genome scans assess the significance of only the main effects of potential disease susceptibility loci. Nevertheless, there is a growing realisation that, when analysing data for complex disease, we need to take into account the possibility of epistasis. This is becoming of increasing importance as SNP (and haplotype) maps are starting to be used in the search for disease susceptibility genes underpinning complex human disease. Statistical methods have already been shown to be computationally tractable. The main difficulty for analysing human complex disease data will be getting large enough samples. So, although a gene's real effect may be too small to detect given the usual sample sizes, it might still be critical, in its interaction on a biological pathway with a second gene.

Most importantly, having found statistical evidence for an interaction, it is imperative to investigate whether this interaction has biological meaning. Unfortunately the degree to which statistical modelling can elucidate the underpinning biological mechanism appears to be limited. A better understanding of the underlying genomic network/s, and the pathways to disease is of vital importance to resolve these fundamental modelling issues.

REFERENCES:

- [1] Skipper, R. A., 2002, The persistence of the R. A. Fisher – Sewall Wright controversy, *Biol. & Phil.*, 17, 341.
- [2] Bateson, W., 1907, Facts limiting the theory of heredity, *Science* 26, 649.
- [3] Wilson, S. R., 2004, Epistasis, *Nature Encyclopedia of the Human Genome*, D. N. Cooper (Ed.), Nature Publishing Group, London, vol. 2, 317.
- [4] Cordell, H. J., 2002, Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans, *Hum. Mol. Genet.*, 11, 2463.
- [5] Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance, *Trans. R. Soc. Edin.*, 52, 1127.
- [6] Cordell, H. J., Todd, J. A., Bennett, S. T., Kawagushi, Y. and Farrell, M., 1985, Two-locus maximum lod score analysis of a multifactorial trait:

- joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes, *Am. J. Hum. Genet.*, 57, 920.
- [7] Cox, N. J., Frigge, M., Nicolae, D. L., Concannon P., Hanis C. L., Bell G. I. and Kong A., 1999, Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nat. Genet.*, 21, 213.
- [8] Cho J. H., Nicolae D. L., Gold L. H., Fields C. T., LaBuda M. C., Rohal P. M., Pickles M. R., Qin L., Fu Y., Mann J. S., Kirschner B. S., Jabs E. W., Weber J., Hanauer S. B., Bayless T. M. and Brant S. R., 1998, Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1, *Proc. Natl Acad. Sci. USA*, 95, 7502.
- [9] Xu J., Langefeld C. D., Zheng S. L., Gillanders E. M., Chang B. L., Isaacs S. D., Williams A. H., Wiley K. E., Dimitrov L., Meyers D. A., Walsh P. C., Trent J. M. and Isaacs W. B., 2004, Interaction effect of PTEN and CDKN1B chromosomal regions on prostate cancer linkage, *Hum. Genet.*, 115, 255.
- [10] Aston C. E., Ralph D. A., Lalo D. P., Manjeshwar S., Gramling B. A., DeFreese D. C., West A. D., Branam D. E., Thompson L. F., Craft M. A., Mitchell D. S., Shimasaki C. D., Mulvihill J. J. and Jupe E. R., 2005, Oligogenic combinations associated with breast cancer risk in women under 53 years of age, *Hum. Genet.*, 116, 208.
- [11] Schork, N. J., Boehnke, M., Terwilliger, J. D. and Ott, J., 1993, Two-trait-locus-linkage analysis: A powerful strategy for mapping complex linkage genetic traits, *Am. J. Hum. Genet.*, 55, 856.
- [12] Maindonald, J. and Wilson, S. R., 2004, Evaluation of genetic and environmental interactions in complex disease, Second RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotypes, Carnegie-Mellon University, 6pp.
- [13] Knapp, M., Seuchter, S. A. and Baur, M. P., 1994, Two-locus disease models with two marker loci: The power of affected-sib-pair- tests, *Am. J. Hum. Genet.*, 55, 1030.
- [14] Dizier, M. H., Babron, M. C. and Clerget-Darpoux, F., 1994, Interactive effect of two candidate genes in a disease: Extension of the marker-association-segregation χ^2 method, *Am. J. Hum. Genet.*, 55, 1042.
- [15] Zinn-Justin, A. and Abel, L., 1998, Two-locus developments of the weighted pairwise correlation method for linkage analysis, *Genet. Epidemiol.*, 15, 491.
- [16] Tang, H-K. and Siegmund, D., 2002, Mapping multiple genes for quantitative or complex traits, *Genet. Epidemiol.*, 22, 313.

- [17] Neuman, R. J. and Rice, J. P., 1992, Two-locus models of diseases, *Genet. Epidemiol.*, 9, 347.
- [18] Li, W. and Reich, J., 2000, A complete enumeration and classification of two-locus disease models, *Hum. Hered.*, 50, 334.
- [19] Schaid, D. J., 1996, General score tests for associations of genetic markers with disease using cases and their parents, *Genet. Epidemiol.*, 13, 423.
- [20] Wilson, S. R., 2001, Epistasis and its possible effects on transmission disequilibrium tests, *Ann. Hum. Genet.*, 62, 565.
- [21] Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. and Moore, J. H., 2001, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *Am. J. Hum. Genet.*, 69, 138.
- [22] Culverhouse, R., Suarez, B. K., Lin, J. and Reich, T., 2002, A perspective on epistasis: limits of models displaying no main effect, *Am. J. Hum. Genet.*, 70, 461.
- [23] Tiwari, H.K. and Elston, R.C., 1997, Deriving components of genetic variance for multilocus models, *Genet. Epidemiol.*, 14, 1131.
- [24] Marchini, J., Donnelly, P. and Cardon, L. R., 2005, Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nat. Genet.*, 37, 413.
- [25] Carlborg, Ö., and Haley, C. S., 2004, Epistasis: too often neglected in complex trait studies? *Nat. Genet. Reviews*, 5, 618.
- [26] Kroymann, J. and Mitchell-Olds, T., 2005, Epistasis and balanced polymorphism influencing complex trait variation, *Nature*, 435, 95.
- [27] Brem, R. B., Storey, J. D., Whittle, J. and Kruglyak, L., 2005, Genetic interactions between polymorphisms that affect gene expression in yeast, *Nature*, 436, 701.