Supplementary Text S1: Different models of interaction

The regression model described in Box 1 is quite general, encompassing a number of different specific cases. Suppose we consider a model of recessive effects (on the log-odds scale) at each of two diallelic interacting loci, so that the binary factors x_B and x_C correspond to indicators of homozygosity for the risk-modifying allele at each locus. The expected log-odds of disease implied by the regression formulation, given an individual's two-locus genotype combination, are shown below:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	α	α	$\alpha+\gamma$
Locus B	b/B	α	α	$\alpha+\gamma$
	B/B	$\alpha+\beta$	$\alpha+\beta$	$\alpha+\beta+\gamma+i$

If, instead, we consider a dominant model, whereby a single allele at each locus is sufficient to modify disease risk, we obtain the expected log-odds:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	α	$\alpha+\gamma$	$\alpha+\gamma$
Locus B	b/B	$\alpha+\beta$	$\alpha+\beta+\gamma+i$	$\alpha+\beta+\gamma+i$
	B/B	$\alpha+\beta$	$\alpha+\beta+\gamma+i$	$\alpha+\beta+\gamma+i$

The actual value of the expected log-odds in each genotype category will depend on the values of the regression parameters α, β, γ and i. For example, under the recessive model, if these parameters took values $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 1$ and $i = 3$, we would obtain log-odds values:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0.5	0.5	1.5
Locus B	b/B	0.5	0.5	1.5
	B/B	1	1	5

The penetrance values (probabilities of getting disease) corresponding to this model (i.e. the values of p rather than of $\ln[p/(1-p)]$) may be calculated using the identity $p = \frac{\exp(\ln[p/(1-p)])}{1+\exp(\ln[p/(1-p)])}$, and are:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0.62	0.62	0.82
Locus B	b/B	0.62	0.62	0.82
	B/B	0.73	0.73	0.99

Note that models with interaction effects on one scale (e.g. the penetrance scale) may correspond to models with no interaction effects on another scale (e.g. the log-odds scale). For example, if under the recessive model on the log-odds scale the regression parameters took values $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 1$ and $i = 0$, we would obtain log-odds values:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0.5	0.5	1.5
Locus B	b/B	0.5	0.5	1.5
	B/B	1	1	2

Here, possession of the risk genotype B/B adds a unit of 0.5 to the log-odds while posession of the risk genotype C/C adds a unit of 1.0 to the log-odds, with no additional (interaction) term required for possession of risk genotypes at both loci. The penetrance values corresponding to this model are:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0.62	0.62	0.82
Locus B	b/B	0.62	0.62	0.82
	B/B	0.73	0.73	0.88

Here, possession of the risk genotype B/B adds a unit of 0.11 to the penetrance while posession of the risk genotype C/C adds a unit of 0.20. However, *subtraction* of an additional -0.05 (i.e. an interaction term) is required when both risk genotypes (B/B and C/C) are possessed. This example illustrates the well-known fact that statistical interaction effects are affected by changes of scale [1]: essentially the regression parameters, including interaction terms, are defined relative to some particular scale of interest. This phenomenon has led to some confusion in terminology [2] concerning whether interaction effects represent departure from a linear (i.e. additive) model or from a multiplicative model, with respect to the main effects of the two loci. A model that is additive on the log-odds scale will be equivalent to a model that is multiplicative on the odds scale, and so departure from either of these models may be considered as equivalent. However, this departure would not be equivalent to departure from multiplicativity on the *original* log-odds scale.

A more general 'genotype' model for the effects of two loci allows for different parameters to represent the effects of having a single copy (i.e. being heterozygous) or two copies (i.e. being homozygous) of a risk-modifying allele, as shown below:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	α	$\alpha+\gamma_1$	$\alpha+\gamma_2$
Locus B	b/B	$\alpha+\beta_1$	$\alpha+\beta_1+\gamma_1+i_{11}$	$\alpha+\beta_1+\gamma_2+i_{12}$
	B/B	$\alpha+\beta_2$	$\alpha+\beta_2+\gamma_1+i_{21}$	$\alpha+\beta_2+\gamma_2+i_{22}$

This model includes nine different parameters: a parameter α that represents the 'baseline' log-odds for an individual who has genotypes b/b and c/c, parameters β_1 and β_2 representing the effects of replacing one or both alleles at locus B with the modifying allele B, parameters γ_1 and γ_2 representing the effects of replacing one or both alleles at locus C with the modifying allele C and four interaction parameters i_{11}, i_{12}, i_{21}, and i_{22}. This is known statistically as a 'saturated' model, which means that it is fully parameterized: nine two-locus genotype categories are modelled by nine parameters, and so these parameters may be chosen (estimated) to fit the observed nine two-locus penetrances or log-odds values precisely. No other model exists that can fit the observed penetrances any better. All other models can be considered as sub-models of ('nested' in) this most general model. Although the saturated model provides the best possible fit to the data, it includes many parameters. In statistical terms, we are usually interested in determining whether a model with fewer parameters can fit the data 'almost as well'. The 4 degree of freedom (df) test of interaction ($i_{11} = i_{12} = i_{21} = i_{22} = 0$) tests whether the interaction terms are required at all. We may also make parameter restrictions to the interaction model to generate fewer df (while retaining one or more interaction parameters) and thus increase power. The recessive and dominant models correspond to models in which certain parameters are set equal either to zero or to each other. An alternative is to assume alleles act additively within a locus, which corresponds to assuming

$\beta_2 = 2\beta_1$, $\gamma_2 = 2\gamma_1$, $i_{12} = i_{21} = 2i_{11}$ and $i_{22} = 4i_{11}$. This restriction converts the nine-parameter 'genotype' model into a four parameter 'allelic' model, $\ln[p/(1-p)] = \alpha + \beta_1 x_B + \gamma_1 x_C + i_{11} x_B x_C$, where x_B and x_C are variables taking values (0,1,2) according to the number of risk alleles at locus B and C respectively. This model contains a single interaction parameter i_{11} that may be freely estimated; a modified version of this model, that makes further restrictions on the relative magnitudes of β_1, γ_1 and i_{11}, has also been proposed [3].

References

[1] Frankel, W. N. and Schork, N. J. (1996). Who's afraid of epistasis? Nat Genet *14*, 371–373.

[2] Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Molec Genet *11*, 2463–2468.

[3] Wang, K. (2008). Genetic association tests in the presence of epistasis or gene-environment interaction. Genet Epidemiol *32*, 606–614.