Contribution of gene-gene interaction to genetic variation and its utilisation by selection

Asko Mäki-Tanila[1] **and William G. Hill**[2]
[1]MTT Agrifood Research Finland, [2]University of Edinburgh

ABSTRACT: We review the contributions of epistasis to the genetic variation expected in segregating populations, utilising models with arbitrary specified genotypic values or others in which there is an underlying additive scale but phenotype is non-linearly related to it, e.g. a multiplicative or threshold model. We show that, even when there is substantial epistasis, rather small amounts of epistatic variance are likely to be generated. With multiple loci contributing to the trait, the proportion of epistatic variance does not generally increase because all the interactions also contribute to main effects and hence the additive variance. Utilisation of even additive × additive variance with an appropriate relationship matrix is unlikely to be effective because the contributions are small and are not retained over generations. Incorporation of genomic data has more promise if it can be focussed on tightly linked regions that can be transmitted across generations, but this will be successful only if such regions contribute substantially to the variance, and such evidence is lacking. At this stage we think selection efforts will be, and should be, focussed on the additive component.
Keywords: epistasis; additive variance; selection; genomics; linkage

Introduction

The genetic comparison of animals is based on their own performance and that of animals sharing genetic factors with them. Their expected genetic similarity is deduced from pedigree information and also now directly using a large number of molecular genetic markers over the genome (genomic breeding values). Genetic improvement programs in purebred populations have to date been concentrated on utilization of additive effects because these are transmissible to offspring and subsequent descendents, but there has been considerable recent interest in identifying epistatic effects among loci. This has been stimulated by searches for causes of the missing heritability of quantitative traits from GWAS studies (Manolio et al. (2009)), and by findings of substantial epistasis in analysis of experimental data (e.g. Bloom et al. (2013), see reviews by Nelson et al. (2013), Mackay (2014)) and recently in segregating populations (Hemani et al (2014)). Even so, the proportion of the genetic variance explained by epistatic terms is typically small relative to the additive (Huang et al. (2013), Hemani et al. (2014)). These and gene functional studies (see review by Phillips (2008)) have generated interest in including interaction effects in genome-wide analyses within populations, including animal breeding stocks. Classical pedigree based improvement programs cannot fully utilize epistatic effects even if present, but new methodology using large scale analyses in collaborative data collection schemes as in human populations (e.g. Hemani et al. 2014) may provide some of the relevant information to do so .

This paper is focussed on epistasis and epistatic variances in the animal breeding context. In improvement of crossbreds there has, of course, also been interest in utilizing heterosis and dominance variance (within gene interactions), but we do not pursue this or utilization of epistatic dominance effects here. We first address the question of how much epistatic variance is to be expected for quantitative traits under different models and how these fit expectation. We extend previous results of Hill et al. (2008), in an attempt to resolve experimental findings of epistasis but not of much epistatic variance, in order then to ask 'should we bother to pursue epistasis?' The particular emphasis is on multiple loci, addressing the point that potential numbers of two, three, …, locus interaction terms rise as n^2, n^3, etc. for n loci, and so may dominate. We note that epistasis can arise from non-linear relationships between an additive underlying genotype and observed phenotype, e.g. for a multiplicative or threshold trait. We also briefly consider the impact of linkage disequilibrium, relevant to *cis*-acting epistatic effects. We then review what the opportunities are for utilising epistasis with pedigree and genomic based methodology. This is, at present, somewhat speculative.

Analysis

The genetic variation is assumed to be caused by n biallelic loci, with all in linkage equilibrium. The allelic effects are assumed to be additive within loci and to generate only additive × additive interactions involving 2, 3, …, n loci. The frequency of the increasing allele B_i at locus i is p_i. For a pair of such loci, the genotypic values are

	b_2b_2	B_2b_2	B_2B_2
b_1b_1	0	a_2	$2a_2$
B_1b_1	a_1	$a_1 + a_2 + [aa]_{12}$	$a_1 + 2a_2 + 2[aa]_{12}$
B_1B_1	$2a_1$	$2a_1 + a_2 + 2[aa]_{12}$	$2a_1 + 2a_2 + 4[aa]_{12}$

For three loci, for example, the genotypic value of B_1b_1 B_2B_2 B_3B_3 is

$$a_1 + 2a_2 + 2a_3 + 2[aa]_{12} + 2[aa]_{13} + 4[aa]_{23} + 4[aaa]_{123}.$$

We restrict subsequent formulae to three loci for simplicity, but extension to more is straightforward. The population mean is

$$\mu = 2p_1a_1 + 2p_2a_2 + 2p_3a_3 + 4p_1p_2[aa]_{12} + 4p_1p_3[aa]_{13} + 4p_2p_3[aa]_{23} + 8p_1p_2p_3[aaa]_{123}.$$

Average effects of gene substitution can be obtained using Kojima's (1959) method. For example:

$$\alpha_1 = \tfrac{1}{2}\partial\mu/\partial p_1 = a_1 + 2p_2[aa]_{12} + 2p_3[aa]_{13} + 4p_2p_3[aaa]_{123} \quad (1)$$

and $V_A = \Sigma_i H_i \alpha_i^2$ where the heterozygosity at locus i is $H_i = 2p_i(1 - p_i)$. Similarly, for interaction effects,

$$[\alpha\alpha]_{12} = \tfrac{1}{4}\partial^2\mu/\partial p_1\partial p_2 = [aa]_{12} + 2p_3[aaa]_{123}$$

$$V_{AA} = H_1H_2[\alpha\alpha]_{12}^2 + H_1H_3[\alpha\alpha]_{13}^2 + H_2H_3[\alpha\alpha]_{23}^2$$

$$[\alpha\alpha\alpha]_{123} = (1/8)\partial^3\mu/\partial p_1\partial p_2\partial p_3 = [aaa]_{123}$$

$$V_{AAA} = H_1H_2H_3[\alpha\alpha\alpha]_{123}^2.$$

The variances depend on the number of loci, allele frequencies, effects and interactions. For illustration, expected variances are computed for two cases (Figure 1), one where $p_i = 0.5$ for all loci, when heterozygosity is maximised, the other for a population where the allele frequency distribution over loci follows that expected in a finite population with selectively neutral mutations (U shape distribution, i.e. $f(p) \propto 1/[p(1-p)]$ assuming a diploid population of size $N = 100$ (when the expected heterozygosity of segregating loci is $E(H) \sim 0.109$. Positive (synergistic) and negative (antagonistic) interaction are considered.

There are two critical points. Firstly, all the gene interaction effects for any locus enter its average effects (1) and consequently contribute to the additive variance; those for epistatic effects include only interactions. Thus for n loci V_A comprises $n \times 4^{n-1}$ terms, V_{AA} comprises $\frac{1}{2}n(n-1) \times 4^{n-2}$ terms and so on. It is therefore not surprising that, with many loci, the additive variance comprises most of the genotypic variance. Secondly, while V_A is a function of heterozygosity at individual loci, those for two (three) locus epistatic interactions depend on products of heterozygosity at two (three) loci. This further reduces them compared to V_A, and higher order epistatic variances are correspondingly smaller too.

The most substantial epistatic variance is likely for antagonistic interaction because these reduce the average effects and hence V_A, as seen in Figure 1, but such cancellation of effects over multiple loci becomes less likely. The impact of the heterozygosity is seen more strongly when the gene frequencies depart from one half as for the U shaped distribution.

Predominance of additive variance is seen in Figure 1, even though the interaction effects (i.e. [*aa*] and [*aaa*] terms) are of similar magnitude to the gene effects *a* at individual loci. It seems reasonable to assume that as the number of loci (n) influencing any trait increases, effects at individual loci are likely to decline roughly in proportion to $1/\sqrt{n}$ (i.e. variance as $1/n$), and similarly two locus epistatic effects $\propto 1/n$. Consequently the relative contribution of the epistatic variance is likely to be smaller than in Figure 1.

Thus we conclude that the finding of small amounts of epistatic variance in segregating populations, even with multiple epistatic loci and substantial epistasis, is exactly what would be expected on simple theoretical grounds (Mäki-Tanila and Hill, in prep.).

Non-linear relationship between genotype and phenotype. Epistatic variance may also arise, even with infinitesimal additive effects on an underlying variable, if there is a non-linear relationship between the underlying genotypic or phenotypic value (x), and observed phenotypic

Table 1. Partition of genetic variance for the observed phenotypes in the multiplicative and threshold multilocus models when there is underlying additive genetic variation with allele frequency 0.5 (VR represents all other components, including dominance variance).

			Multiplicative model			Threshold model			
h^2	loci	CV_P	V_A/V_G	V_{AA}/V_G	V_R/V_G	P($x>T$)	V_A/V_G	V_{AA}/V_G	V_R/V_G
0.1	5	0.3	0.996	0.004	0.00045	0.30	0.986	0.0114	0.0021
	∞		0.996	0.004	0.00001		0.985	0.0137	0.0010
	5	0.5	0.990	0.009	0.00116	0.05	0.887	0.0975	0.0155
	∞		0.989	0.011	0.00008		0.877	0.1176	0.0048
0.9	5	0.3	0.965	0.030	0.00451	0.30	0.749	0.0860	0.1644
	∞		0.962	0.037	0.00098		0.736	0.0928	0.1710
	5	0.5	0.913	0.074	0.01381	0.05	0.317	0.3520	0.3312
	∞		0.903	0.091	0.00639		0.326	0.3952	0.2790

value (y), such as with a multiplicative (cf. Dillham and Foulley 1998), optimum (Haldane 1954) or threshold model (cf. Dempster and Lerner 1950). We consider how the degree of non-linearity is reflected in the partition of the genetic variation of the observed trait into additive and non-additive components, assuming that the underlying genetic variation is additive with normally distributed environmental deviations. For the multiplicative case, $y = \exp\{kx\}$ where k is a scaling factor. For an optimal trait, the fitness is assumed to have quadratic or nor-optimal form, $y = 1 - (x - x_{opt})^2/k$. For a threshold trait the observed phenotype is the probability $y > T$, for a threshold value T. Variance components are obtained by differentiating the mean of y with respect to allele frequencies (as eq. 1).

The proportion of epistatic variance increases in the multiplicative model as the coefficient of genotypic variation ($CV_G = \sqrt{V_G}/\mu$) becomes higher (Table 1), in the threshold model as the mean probability departs further from 0.5 (Table 1), and in the optimum model as the population mean becomes closer to the optimum. In each case this is when the relation between the underlying genotype and observed phenotypes is most highly non-linear. Most of the epistatic variance is contributed by V_{AA}, as higher derivatives are smaller for such non-linear continuous functions. Only for the threshold model and with very extreme probabilities does V_{AAA} contribute much. More epistatic variance results with higher heritability values (h^2) on the underlying scale, when the environmental variance does not dilute the non-linear relationship. Although the underlying variable is additive, dominance variance and interactions are also produced, but these are small unless only a few loci of large effect contribute to the genetic variance in x.

As the number of loci becomes large, the additive variance on the observed scale can be deduced from the covariance between observed phenotype and underlying genotype. For example, in the multiplicative model $\mathrm{cov}(x,\exp\{kx\}) = \mu h^2 V_x$, where h^2 and V_x are on the underlying scale,

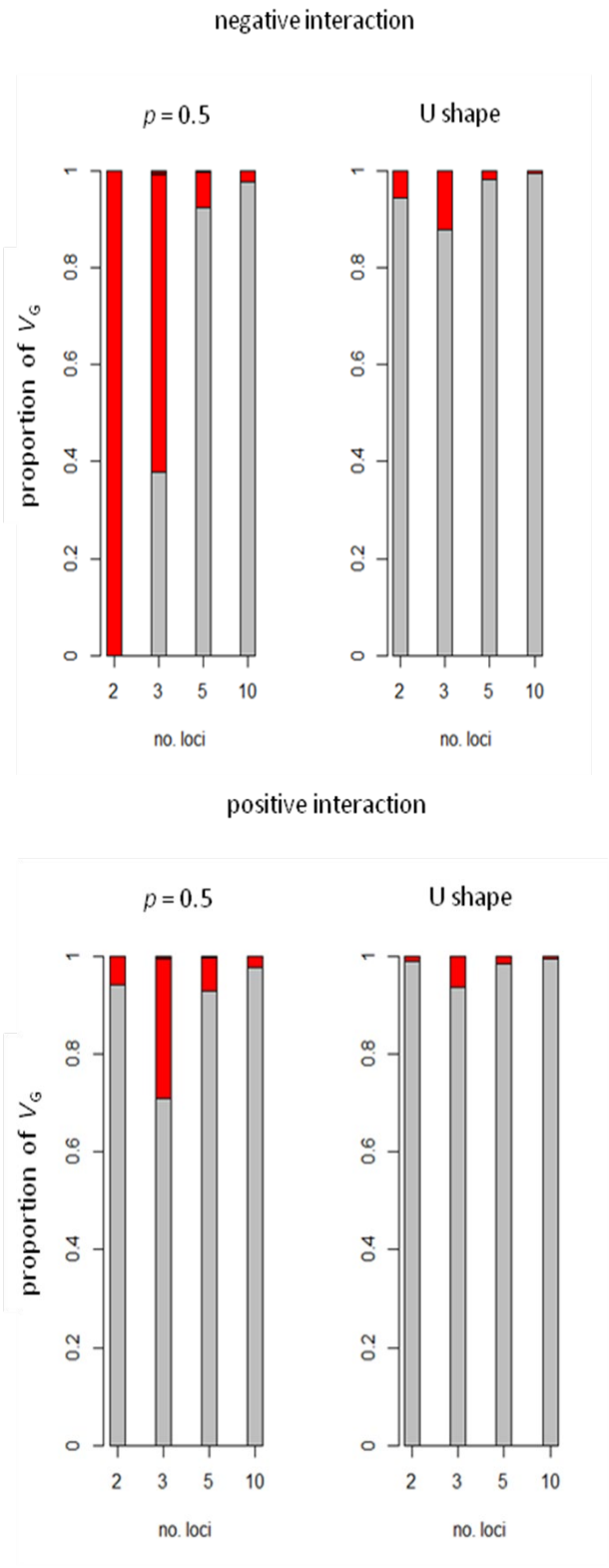
$$V_A/V_G = \ln(1 + CV_G^2)/CV_G^2$$

on the observed scale (Cockerham, 1959). Hence it can be shown that

$$V_A/V_G = h^2 \ln(1+ CV_P^2)/[(CV_P^2 + 1)^{h2} - 1]$$

on the observed scale. For the threshold and optimum models the additive variance is obtained by differentiation, showing analogy between Robertson's method (in Dempster & Lerner (1949)) and Kojima's (1959). In practical animal breeding, the models can be linearised by log-transformation in the multiplicative model, using different kinds of liability model for threshold traits and concentrating on linear local change in the quadratic curve of the optimum model.

Figure 1. Partition of genetic variance in the multilocus model with positive ([*aa*] = [*aaa*] = *a*) and negative ([*aa*] = [*aaa*] = *a*) two and three locus interaction (V_A grey, V_{AA} red, V_{AAA} dark red) for allele frequency 0.5 and for U shaped allele frequency distribution.



Influence of linkage disequilibrium among very tightly linked loci. In the presence of linkage disequilibrium (LD), effects at different loci are not orthogonal so it is not possible to partition variation between contributions of individual loci, except arbitrarily, e.g. fitting locus A and then B after A. Hence we consider contributions from pairs of linked loci simultaneously, and just estimable quantities such as the genotypic variance and the covariances of close relatives. More general results were given by Gallais (1974).

We consider contributions from pairs of loci with the same model of genotypic values, with only additive and additive × additive effects. The LD coefficient is D and haplotype frequencies are, for example, freq(B_1B_2) = p_1p_2 + D. There is random mating, with Hardy-Weinberg applying to genotype frequencies in terms of haplotype frequencies. To simplify formulae yet exemplify the main points, we assume linkage is sufficiently tight that there is negligible recombination over a single generation, i.e. only non-recombinant gametes are produced. This is justified because substantial LD between loci can have arisen in finite populations, even with some selection, only if linkage is very tight (order 1/N) or selection very strong. As we consider only first generation relatives, i.e. sibs and offspring, we are ignoring only one generation's recombination.

The overall mean is

$$\mu = 2p_1a_1 + 2p_2a_2 + (4p_1p_2 + 2D)[aa]_{12}.$$

After some messy algebra, the genotypic variance can be shown to be

$$V_G = 2p_1(1 - p_1)a_1^2 + 2p_2(1 - p_2)a_2^2 + 4Da_1a_2 + 4\{2p_1(1 - p_1)p_2 + D\}a_1[aa]_{12} + 4\{2p_1p_2(1 - p_2) + D\}a_2[aa]_{12} + \{4p_1p_2(1 + p_1 + p_2 - 3p_1p_2) + 2D(1 + 2p_1 + 2p_2 - 4p_1p_2)\} [aa]_{12}^2$$

Note that, in the absence of epistasis, i.e. $[aa]_{12}$ = 0, $V_A = V_G$ and the variance comprises contributions due to variances at individual loci and the term $4Da_1a_2$ from the covariances between loci. As it occurs in subsequent formulae, let

$$E = \{p_1(1 - p_1)p_2(1 - p_2) + D^2\}[aa]_{12}^2 = p_1(1 - p_1)p_2(1 - p_2)(1 + r^2)[aa]_{12}^2,$$

where r is the correlation of gene frequencies.

The covariance of parent and offspring (cov_{OP}) in the absence of recombination, such that individuals transmit their constituent haplotypes entire, is equal to twice the half-sib covariance. These turn out to be

$$cov_{OP} = 2cov_{HS} = \tfrac{1}{2}V_G - E.$$

In the presence of epistasis this does not accord with familiar expressions, $cov_{OP} = \tfrac{1}{2}V_A + \tfrac{1}{4}V_{AA}$ and $cov_{HS} = \tfrac{1}{4}V_A + (1/16)V_{AA}$, but these apply with free recombination when alleles at different loci transmit independently. With negligible recombination, the covariance of full sibs comes from 50% sharing one haplotype in common, contributing cov_{HS}, and 25% sharing both, contributing $\tfrac{1}{4}V_G$. Hence

$$cov_{FS} = \tfrac{1}{4}V_G + cov_{HS} = \tfrac{1}{2}V_G - \tfrac{1}{2}E.$$

Hence the covariance of full sibs within half sibs is $\tfrac{1}{4}V_G$ and the sire × dam interaction in a diallel analysis is

$$V_{S\times D} = cov_{FS} - 2cov_{HS} = \tfrac{1}{2}E.$$

These results apply, of course, for any arbitrary pair of loci under the same assumptions.

The proportion of the variation contributed by the utilizable ('additive') components, e.g. cov_{OP}, is large, compared to e.g. $V_{S\times D}$ which involves only the epistatic term E. This is because E is a function of products of heterozygosity (potentially doubled if $r^2 = 1$), in contrast to the 'additive' components which are functions of heterozygosities or third order terms in gene frequency. We see therefore that the component of epistatic variance that influences the different types of covariance is a small proportion of V_G. Enumerated examples, not shown, support this conclusion.

Possibilities for utilizing epistasis

Ideally we should give answers to questions such as: what kind of gains are there from including epistatic deviations among selection criteria; can the epistatic variance be utilized using classical selection using phenotypic and pedigree information; and what potential is there by also incorporating genomic information?

Covariances among relatives and selection. Formulae for contributions of epistatic variance to covariances among relatives date back to Fisher (1918) and to the work of Cockerham (1954), Kempthorne (1954) and subsequently Bulmer (1980). Even though amounts of epistatic variance are small, we should understand how it might be utilized in breeding programs.

The covariance among relatives due to interactions involving additive effects from unlinked loci is given by powers of the relationship, as noted above, e.g.

$$cov_{FS} = cov_{OP} = \tfrac{1}{2}V_A + \tfrac{1}{4}V_{AA} + (1/8)V_{AAA} + \ldots,$$

and therefore the epistatic terms contribute proportionately much less to covariances among more distant relatives, e.g. for grandoffspring and grandparent

$$cov_{GOGP} = \tfrac{1}{4}V_A + (1/16)V_{AA} + (1/32)V_{AAA} + \ldots$$

Thus the regression of offspring on parental mean phenotype equals $(V_A + \tfrac{1}{2}V_{AA} + \ldots)/V_P$ and that on grandparental mean phenotype is $(V_A + \tfrac{1}{4}V_{AA} + \ldots)/V_P$.

Half the gain from V_{AA} obtained in the progeny is lost in the grandprogeny, so gains from epistasis are not cumulative, (Griffing (1960); Bulmer (1980); Crow (2008)).

Genetic evaluations should contain all possible factors affecting the variation. When relevant factors are omitted, the accuracy is reduced, e.g. in breeding value estimation the neglect of non-additive effects would result in losses dependent on their variance. Existence of epistatic variance also causes a bias in estimates of the additive variance from close relatives. Non-additive effects have been largely ignored because of the highly involved machinery and extra computing time required for their estimation. The practical problem is to have sufficient data to distinguish between different models and obtain adequate estimates of effects. We should also note that the breeding values and epistatic effects are to a large extent confounded, e.g. the elements of the relationship matrix for A×A terms are squares of those of the additive relationship matrix. The theory shows that even ubiquitous gene interactions in a multi-locus quantitative genetic system is expected to produce only small amounts of epistatic variance within populations in addition to the additive variance. Even with very large data sets, estimates of the contributions of epistasis in outbred populations have (so far) been in accordance with these expectations.

Epistatic contributions to covariances among relatives are influenced by linkage (Cockerham 1956; Lynch and Walsh 1998) and, importantly, selected haplotypes are retained across generations. For a pair of loci with recombination fraction c, the expression for regression on parental mean now becomes $(V_A + (1 - c)V_{AA})/V_P$ and that on grandparental mean phenotype is $(V_A + (1 - c)^2V_{AA})/V_P$. Coefficients of V_{AAA} depend on retention of the corresponding full three locus haplotype. Similarly covariances among sibs are different functions of c and so are not simple multiples of those between generations.

Across the genome as a whole, the mean recombination fraction is very close to one-half because most pairs of loci lie on different chromosomes or far apart on the same chromosome. Therefore putting selection effort on utilizing epistasis would seem to be wasted unless linkage can be employed. Also, in view of the limited amount of epistatic variance, it is most likely to be useable if concentrated among linked sites on which selection can be focused. This implies that conventional selection based on overall pedigree is unlikely to be effective in utilizing epistasis, as many, dating back to Lush, have appreciated. In any case, weighting on average effects in pedigree or genomic analyses automatically also puts weight onto epistatic sites. Does incorporation of genomic information change this picture?

Genomic tools. There are very few cases where a single gene explains a substantial fraction of the variation in a metric or disease trait. It is even less likely that pairs or trios of loci with a major contribution will be found. In genome-wide analyses, the number of effects to be estimated is the square of that for individual loci, so with many thousands of markers very stringent test criteria have to be used and therefore the power is very low. It has become obvious that GWAS cannot harvest all the existing genetic variation, in particular that due to rare alleles is often undetected. Such problems are greater in considering interaction effects, where estimation also depends on LD of two pairs of markers and trait genes; and interaction effects are likely to be of smaller order than main effects with multiple loci. Hence detection of many epistatic effects in livestock populations is unlikely, although with sufficient resources not perhaps impossible.

Machine learning methods rather than linear models have been suggested as a way to fit single locus and epistatic components of any order (Gianola and de los Campos (2008), Long et al. (2008)) for example in prediction of survival of progeny of individual sires. Whilst such a method can identify the best animals to breed offspring, it does not identify those likely to have the best grandoffspring and beyond because it does not reflect the Mendelian transition process whereby genotypes are re-assorted each generation and only the additive components persist.

We consider different scenarios, informed by GWAS analysis in the population. If a GWAS analysis, for example, reveals no substantial specific pair-wise interactions, even though there is evidence of substantial epistatic variance from pedigree studies, it seems unlikely to be worthwhile to try to utilize it rather than focusing on additive effects at individual loci, whether utilizing pedigree or genomic relationship information. It could be done using a two site genomic relationship matrix, basically a square of elements at individual sites, but this seems unlikely to add much to the relationship matrix *per se* but add greatly to computation.

A second scenario is where a GWAS or similar analysis reveals substantial interactions among two or more sites for traits of interest, but these sites are essentially unlinked (say $c > 0.4$), i.e. *trans* effects. In principle, markers around both sites (and sets of sites) could be fitted in addition to the overall genomic relationship matrix or Bayesian scenario fitted using individual gene effects. The benefits from doing so might be small, as recombination would imply that only one quarter of the favored haplotypes are retained and devoting selection to them rather than to additive effects throughout the genome would surely not be justified.

The third and more interesting scenario is where a GWAS or similar analysis reveals regions of large effect on the trait and where fitting individual SNPs within regions identifies *cis* epistatic effects, also of course implying that LD between the trait loci is not complete. It would then be feasible to fit the markers in the region jointly, perhaps adding it into the Bayesian weighting of individual sites. Details and practicalities have yet to be explored. The critical point is that such epistatic haplotypes would be transmitted and the use of the genomic data enables selection among them. For example, it might be useful in

selection to eliminate an undesirable allele pair, for example causing disease or infertility. Thus we see a narrow window of opportunity, but suspect it is only such.

The selection on linked interacting major genes is worth pursuing although compromises in the overall efficiency of selection follow. As with marker-assisted selection with major genes, care should be attached to the use of promising marker pairs in selection as relying on false positives would impair the accuracy of selection compared to the case of not using them.

Conclusions

In conclusion, epistatic variance is expected to be small on theoretical grounds and recent observations support this. It is also hard to exploit beyond small and temporary gains in selection response and can generally be ignored in breeding programs within populations. Genomic predictions with large marker panels are also likely to gain very little by including interaction effects in the analyses. More research is needed about the obviously rare cases of major genes with substantial interaction and how the variation of such tightly linked genes could be harnessed.

Acknowledgements

AMT acknowledges the receipt of a fellowship from the OECD Co-operative Research Programme: Biological Resource Management for Sustainable Agricultural Systems in 2013 for the visit to Edinburgh.

Literature Cited

Bloom JS, Ehrenreich IM, Loo WT et al. (2013). Nature 494: 234–237.

Bulmer MG (1980). *The mathematical theory of quantitative genetics.* New York: Oxford University Press.

Cockerham CC (1954). Genetics 39: 859–882.

Cockerham CC (1956). Genetics 41: 138-41.

Crow JF (2008). J. Genet. 87: 349–353.

Dempster ER, Lerner IM (1950). Genetics 35: 212-236.

Dillmann C, Foulley JL (1998). Gen. Sel. Evol 30: 543-564.

Fisher RA (1918). Trans. Roy. Soc. Edin. 52: 399–433.

Gallais A (1974). Biometrics 30: 429-446.

Gianola D, de los Campos G (2008). Genet. Res. 90: 525-540

Griffing B (1960). Austr. J. Biol. Sci. 13:307-343.

Haldane JBS (1954) Proc. IX Int. Congr. Genet., Caryologia (Suppl.), 480-487.

Hemani G, Shakhbazov K, Westra HJ et al. (2014). Nature doi:10.1038 Published online 26 Feb 201:

Hill WG, Goddard ME, Visscher PM (2008). PLoS Genet 4: e1000008.doi:10.1371/journal.pgen.1000008

Huang W, Richards S, Carbone MA et al. (2012). Proc. Natl. Acad. Sci. USA. 109:.15553-15559.

Kempthorne O (1954). Proc. R. Soc. Lond. B Biol. Sci. 143: 102–113.

Kojima K (1959). Proc. Natl. Acad. Sci. USA 45: 984–989.

Long N, Gianola D, Rosa GJ et al. (2008). J. Anim. Sci. 86: 3358-3366.

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

Mackay TFC (2014). Nat. Rev. Genet. 15, 22–33.

Nelson RM, Pettersson ME, Carlborg Ö. (2013). Trends Genet. 29: 669-676.

Manolio TA, Collins FS, Cox NJ et al. (2009). Nature 461: 747-753.

Phillips PC (2008). Nat. Rev. Genet. 9: 855–867.