Complex Trait Genetics and Gene-to-Phenotype Models

Mark Cooper, Dean Podlich and Oscar S. Smith

Pioneer Hi-Bred International Inc., 7250 N.W. 62nd Avenue, P.O. Box 552, Johnston, Iowa 50131, USA,
Email mark.cooper@pioneer.com

Abstract
A plant breeder has to deal with multiple traits and many of these are genetically complex. The technologies that support plant breeding have progressed to a stage where there are now many options available to the applied breeder for the design of a breeding strategy. However, at this time the efficacies of many of the molecular breeding strategies that have been proposed for complex traits have not been empirically evaluated and compared to progress from conventional selection on phenotype. We seek a theoretical framework to better understand the power of phenotype-based (conventional) and molecular-based plant breeding strategies to change multiple complex traits by selection, and to study their relative strengths and weaknesses. For many traits high throughput technologies for studying DNA sequences have enabled us to move from studying phenotypes to the identification of candidate genomic regions and genes. To complement and focus our gene discovery capabilities we seek appropriate methods to develop gene-to-phenotype (GP) models that will lead to molecular-based strategies that are more efficient than the conventional pedigree-based breeding process. Advances in computer simulation, combined with large experimental data sets, provide the opportunity to consider the genetic architecture of traits on a continuum from simple to complex. We discuss the foundations of a suitable quantitative framework and apply this to examine aspects of response to selection. With this framework we can show that as the complexity of the genetic architecture of traits increases the opportunities for improving on phenotypic selection by molecular-enhanced strategies increase, but in parallel the requirements for development of adequate GP models become more challenging.

Media summary
Effectiveness of improving complex traits using molecular breeding strategies can be modeled using advances in computer simulation, theoretical gene-to-phenotype models and large experimental data sets.

Key Words
Prediction, *E(NK)*, Interaction, Epistasis, G×E, QTL

Introduction
Plant breeders have always been confronted with the problem of predicting the expected phenotypic performance of new individuals with untested gene combinations (new genotypes) with limited information on the gene-to-phenotype (GP) architecture for traits. The pedigree-based breeding strategies used today have emerged from a continual process of testing and refinement by applied breeders. There are opportunities to apply molecular technologies to further refine these breeding strategies. Ultimately it will not be sufficient to demonstrate that we can predict phenotypic variation and the phenotypic changes that result from selection using genetic information, but that this knowledge allows us to improve on the outcomes that are currently being achieved by conventional selection on phenotype alone. To examine the potential of molecular-enhanced breeding strategies to achieve this end, we apply a theoretical framework for GP models that includes important details of the genetic architecture of complex traits, e.g. epistasis, gene-by-environment interactions. With this theoretical framework it is becoming feasible to undertake evaluations of the merits of molecular-enhanced breeding strategies.

While genetics provides the scientific basis for the breeding processes we use today, for the majority of the history of applied breeding the concept of a gene was unknown. Selection was conducted on the phenotypes of individuals in ways that was a mimic of the Darwinian process of natural selection, which was itself an undefined concept prior to 1859. Further, for the majority of the 20th Century the genes underlying quantitative traits were at best theoretical constructs and were not directly viewed or manipulated in breeding programs to bring about changes in phenotype. It was only in approximately the last quarter of the 20th Century that we have observed technological advances that provided the opportunities to study genetic variation at the DNA sequence level. These technologies have enabled us to

26 Sep – 1 Oct 2004, Brisbane, Australia. Published on CDROM. Web site www.cropscience.org.au

go from the phenotype to the gene, and in some cases from the gene to the trait phenotype. Many attempts have been made and are currently underway to construct relevant GP models for traits to assist the plant breeding process. The approaches being used are diverse and in many cases unproven. This is a time of exploration of many novel ideas on how to approach the GP problem.

Quantitative genetic theory, with all of its assumptions, was founded with the goal to understand the genetic basis of the variation for quantitative traits and to use this knowledge to make predictions about the properties of genes in populations of genotypes and the outcomes of artificial selection and evolutionary processes. The assumptions we made in constructing these models appeared reasonable at the time and given the available experimental data. Taking a broad view of the relationship between the predictions from applying this quantitative genetic theory to applied plant breeding, Coors (1999) summarized many of the published recurrent selection studies for the quantitative trait grain yield in corn. The synopsis we can take from Coors' synthesis of published studies strongly suggests that the realized progress from selection for this trait is considerably lower than the predicted response. For most involved in applied breeding this result is not surprising. However, this quantified observation forces us to consider the possible reasons for the discrepancies between the predictions made from classical quantitative genetic theory and the realized responses from applied breeding. We may expect that some of the simplifying assumptions that we routinely make in classical quantitative genetic theory are incorrect and at the core of this discrepancy. However, of the many assumptions made which are more important in determining the gap between predictions and realization of genetic progress? Further, can we use the genetic knowledge of today to construct improved GP models and would these in turn improve our ability to predict the expected outcomes from a breeding program? Despite the significant advances we have made in the range of molecular technologies available to study the genetic architecture of traits, the GP prediction problem still exists today as a major challenge for most of the important traits in plant breeding. If anything, today the magnitude and complexity of the task involved in predicting phenotypic variation based on knowledge of DNA sequence variation is now more obvious to a wider audience than was previously the case. We will discuss some of the issues involved in dealing with this complexity and how these are relevant to the design of conventional and molecular breeding strategies.

Here we describe components of a quantitative framework that can be used to investigate some expected properties of traits under the influence of selection for the continuum of simple to complex traits. Classical quantitative genetics, in combination with linear statistical models, provides the conventional approach to prediction of expected response to selection. The framework we describe here differs from much of the classical theoretical framework of quantitative genetics in that it is implemented and the predictions are obtained through computer simulation rather than through seeking solutions to linear statistical models using approaches from calculus. The motivation for developing and using a simulation framework is the difficulty of extending the classical statistical framework to accommodate recognized properties of complex traits, particularly effects attributed to features such as gene-by-gene interactions (epistasis) and gene-by-environment interactions. The need to accommodate these properties of gene action in many GP models is indicated by experimental investigations into the molecular basis of regulation of gene expression, signal transduction pathways and other features of genetic variation at the DNA sequence level and their influences on phenotypic variation of traits within an organism that is undergoing growth and development within an environmental context.

The use of computer simulation to study problems in genetics is not new (*e.g.* Fraser and Burnell 1970). However, many of the early applications of simulation were based on scaling up the classical simple Mendelian inheritance models, assuming independence of gene effects from the genes at other loci and of environment. Thus, many of the complex interactions we now consider as components of the GP models for some of the important traits were not included as features of the early simulation experiments. This limitation is not unique to the early genetic simulation experiments. Many of the applications of simulation to problems in quantitative genetics today still only consider genes as independent Mendelian factors. Such limited treatment of the continuum of GP models is unnecessary today and is not recommended. Our own work suggests that such approaches will give an overly optimistic and simplistic view of the expected outcomes of conventional and molecular breeding strategies (Cooper and Podlich 2002; Chapman et al 2003; Peccoud et al. 2004). This body of work also gives some indicators of features of the GP relationships that will contribute to the observed discrepancies between predicted and realized genetic gain for quantitative traits.

The key components of the simulation framework considered in this paper are: (1) definition and elaboration of the *E(NK)* model as an organizing framework for studying GP relationships for traits; (2) use of fitness landscape concepts to study the continuum of simple to complex GP models; (3) predicting phenotype from genotype and characterization of GP properties using both classical quantitative genetic models and landscape specific parameters; (4) studying the continuum from simple to complex trait genetics and factors that have a strong influence on the outcomes of directional selection; and (5) some preliminary considerations related to interpretation of short-term and long-term responses to selection.

The E(NK) *model for traits*

The *E(NK)* model we discuss here is an extension of the *NK* gene network model that was introduced and used by Kauffman (1993) to study the behavior of gene networks and their influences on organism development and evolutionary processes. Here we have confined our application of the *E(NK)* model to the study of issues as they are relevant to plant breeding processes. The *E(NK)* model allows for the property that the influence of a gene network on determination of a trait phenotype can differ among environmental conditions. Thus, *E* identifies different environment-types within the context of a defined target population of environments, *N* identifies the different genes and *K* identifies the degree of connection between subsets of the total set of *N* genes, *i.e.* the gene network topology. Thus, in the terminology of quantitative genetics the *E(NK)* model is a finite locus polygenic model that can be defined to include effects of epistasis and gene-by-environment interactions. The parentheses around the *NK* term are used to indicate that the *N* genes can interact in different *K* ways to determine the trait phenotype in different *E* environment-types. To date we have used the *E(NK)* model as an organizing framework for the design of large scale computer simulation experiments that are conducted to investigate both the properties of GP relationships for a continuum of simple to complex genetic models and the power of different plant breeding strategies to achieve response to selection along this complexity continuum.

Kauffman (1993) gave a detailed explanation of the background and specification of the *NK* model. The background to the *E(NK)* genetic model and descriptions of its application to specific traits within a plant breeding context have been given elsewhere (Podlich and Cooper 1998; Cooper and Podlich 2002; Cooper et al. 2002; Peccoud et al. 2004). Here we relate the definition of the *E(NK)* model to the classical finite locus models used in quantitative genetics (*e.g.* Falconer 1960, Falconer and Mackay 1996; Lynch and Walsh 1998). We consider the use of "*hybrid models*", where the genetic component of the model is based on a combination of "*explained*" genetic variation attributed to defined genes (or Quantitative Trait Loci; QTL) and "*unexplained*" background genetic variation. The *explained* component of the genetic variation may be defined as an outcome of investigations into the inheritance of a trait by use of suitable experimental methods, such as inheritance studies using a genetic or sequence based map of the genome.

Without any loss of generality, in this paper we discuss GP models for traits within the context of a typical quantitative trait mapping experiment that can be conducted by a plant breeder. A typical experiment involves testing a sample of genotypes from a reference mapping population in a sample of environments. The phenotype for the *k*th observation on the *i*th genotype in the *j*th environment (P_{ijk}) can be described by the linear equation:

$$P_{ijk} = E_j + G_i + (GE)_{ij} + \varepsilon_{ijk} \quad , \quad (1)$$

where, E_j is an environmental effect, G_i is a genotypic main-effect, $(GE)_{ij}$ is a genotype-by-environment interaction effect, and ε_{ijk} is a residual effect. Linear statistical models, of the form given in equation (1), can be used to analyze means and variances associated with the genotype-environment system based on the experimental sample. It should be recognized that equation (1) is not a GP model; it is a statistical partition of inter-individual genotypic variation. For a GP model we need a specification of the trait performance as an outcome of the combined effects of the *N* genes influencing the trait. If we use γ_n to refer to gene *n*, where $n = 1, \ldots, N$, and $\gamma_{g|n}$ to represent the specific *g* genotypic combinations of alleles for gene *n*, then any individual, represented as a multi-genic genotype G_i, can be considered to have a value in environment E_j from the combined effects of the *N* genes in that environment. We identify the combined multi-genic effects of the *N* genes for genotype *i* in environment *j* by $\Gamma_{(g|N)ij}$, and rewrite equation (1) as a GP model in the form:

$$P_{ijk} = \Gamma_{(g|N)ij} + \varepsilon_{ijk} \quad . \quad (2)$$

We use equation (2) as a compact statement that defines the phenotypic value for a trait measurement *k* on a genotype *i* in an environment *j* as the genotypic value, that is the outcome of the combined effects of the allele combinations for the *N* genes, and a noise parameter ε_{ijk} that can be interpreted as a random environmental and/or measurement error effect. From the above, the *E(NK)* model is an alternative form of the genotypic component of equation (2) that identifies the combined effects of the *N* genes by the *NK* term and the specificity of the resulting *NK* genotype values to an environment by nesting the $(NK)_i$ term for genotype *i* within a defined environment E_j. Thus, we can rewrite the *E(NK)* model in a form similar to equation (2) as:

$$P_{ijk} = E_j(NK)_i + \varepsilon_{ijk} \quad . \quad (3)$$

Thus, equations (2) and (3) define the rudimentary framework for computing the genotypic and phenotypic values for *i* genotypes in *j* environments from knowledge of how the *N* genes act to determine the trait phenotype. More generally we can recognize that the *E(NK)* model is itself a special case of the general expression for the equations of state of a system given by Casti (1997, pp.7-10):

$$y = \Phi_\alpha(u) \quad , \quad (4)$$

where; *y* is the phenotype, Φ is a mathematical relationship expressing the relationships among the observables, α is a parameter vector representing the genetic component of a system and *u* represents environmental conditions. Recognizing this relationship identifies and opens up a wide array of possibilities for studying the GP problem state space that go beyond the conventional framework of quantitative genetics.

Predicting Phenotype from Genotype

If we seek to improve the effectiveness of the conventional plant breeding process by using molecular technologies to design a knowledge-based approach to plant breeding, equations (2) and (3) emphasize that one of the major challenges that we face is how to discover the *N* genes that are important in determining the extant phenotypic variation for a trait and how to understand the various functions of these genes. Presently we rely heavily on our suite of forward and reverse genetics approaches to identify a relevant subset of the *N* genes. Functional genomics technologies can then be applied in combination with appropriately designed genetic experiments to provide the basis for diagnosing the interactions among the genes and the construction of gene network knowledge. A model based on the integration of such knowledge would then provide a starting point for understanding our current capacity to predict changes in the phenotype from directed changes in the genotype. Two approaches we can take to examine the magnitude of this GP prediction problem are: (1) Undertake an extensive experimental program to discover the genes and understand their function and role in determining phenotypic variation for the target traits, an activity that is currently underway in many research groups. Then use this knowledge to make predictions and conduct validation experiments to test the predictions. (2) Use the *E(NK)* model to create a simulated ensemble of many different "plausible" GP models and examine the robustness of alternative breeding strategies across these different GP models. We can combine both approaches by using the available experimental data to place the empirical GP models into the theoretical GP problem space. Within this paper we consider aspects of both approaches.

From the body of quantitative trait mapping literature available today we observe that for every study we will obtain a statistical model that is a partial representation of the genotypic components of the phenotypic variation. Thus, we can rewrite a qualitative version of equation (1) to represent this situation as:

$$P_{ijk} = E_j + G_{Expl} + G_{UnExpl} + (GE)_{Expl} + (GE)_{UnExpl} + \varepsilon_{ijk} \quad , (5)$$

where the subscripts *Expl* and *UnExpl* identify *explained* (or estimated) and *unexplained* components of the G and G×E components of the linear model, respectively. Here, we treat the *explained* and *unexplained* components of the model as independent terms. It is emphasized that this independence is not necessary and is only a subset of the more general case. If the *explained* component interacts with the *unexplained* component we have a much more complex situation than is indicated by equation (5). In the presence of such interactions we may expect to overestimate the level of predictability within the system based on the GP knowledge we have acquired. A tangible example of this situation would be when we move alleles of genes from one population into another and observe different effects of these alleles in these different contexts. Because of space constraints we do not give a formal treatment of the more general case here. Instead below we include some preliminary discussion of its implications for predictability within the genotype-environment system.

We can quantify equation (5) by replacing the *explained* component directly with the current version of the GP model for a trait. For example, if we had identified two QTL for the trait and QTL 1 is constitutive in its behavior and effects across environments whereas QTL 2 is facultative, with effects on the trait phenotype that are specific to the types of environment, using our notation above, but using *Q* for a QTL rather than γ for a gene, we could write equation (5) as:

$$P_{ijk} = E_j + Q_{(g|1)i} + Q_{(g|2)ij} + G_{UnExpl|i} + (GE)_{UnExpl|ij} + \varepsilon_{ijk} \quad , \quad (6)$$

where the estimates of the 2 QTL effects represent an *explained* component. While we can use estimates of QTL effects as a parameterization of the *explained* component of the model, this opens the question of how to appropriately represent the *unexplained* component. A classical approach would be to assume that the *unexplained* component was independent of the *explained* component and can be represented by effects drawn from some underlying distribution, such as the normal distribution. However, if we apply the same general arguments regarding the limitations of the assumptions of independent gene effects given above to the currently *unexplained* component we may consider that representing the unknown component of the model in such a manner may be an overly optimistic representation. An alternative approach that we have considered is to represent the known component by the QTLs and the unknown component by an ensemble of *E(NK)* model parameterizations. Thus, we would rewrite equation (6) as:

$$P_{ijk} = E_j + Q_{(g|1)i} + Q_{(g|2)ij} + E_j(NK)_{ij} + \varepsilon_{ijk} \quad , \quad (7)$$

where we have retained the two QTL from equation (6) as the *explained* genetic component and defined the *unexplained* genetic component as an *E(NK)* model parameterization. A macro-environmental effect E_j and a micro-environmental noise effect ε_{ijk} associated with measurement *k* on genotype *i* in macro-environment *j* are both retained in this model expression. The E_j term could in fact be omitted but we leave it in here for consistency and we define the macro-environments to be the same as the environment-types specified within the $E_j(NK)_{ij}$ component of the model. Equation (7) is an expansion of equation (3) using empirical results from the currently available GP model. Equation (7) can be parameterized by specifying the effects for a subset of the *N* genes defined in equation (3) based on experimental estimates of their effects and specifying the remainder of the *N* genes by sampling effects from some underlying distribution. We refer to equation (7) as a *hybrid model*, where the *explained* component is parameterized by estimates of gene effects obtained from experimental results and the *unexplained* component is stochastically parameterized by drawing effects from a specified distribution of effects. Equations (5) - (7) represent one example of how this approach can be implemented within the framework indicated by equation (3). We can indicate a more general application for QTL models by writing:

$$P_{ijk} = \underset{Expl}{\left[E_j(NK)_{ij} \right]} + \underset{UnExpl}{\left[E_j(NK)_{ij} \right]} + \varepsilon_{ijk} \quad . \quad (8)$$

Equation (8) can be applied to incorporate effects attributed to epistasis in the form of QTL×QTL interactions in the *explained* or *unexplained* component. Further extensions of this framework that we have considered include allowing some of the *K* interactions to occur between some of the *N* genes in the *explained* component and those in the *unexplained* component to simulate the effects of partial

explanation of a gene network or the gene-by-background effects often experienced in plant breeding, where QTL or gene effects are population (cross) specific. We can expect to observe a range of simple to complex situations and a mixture of successes and failures in breeding programs as we attempt to validate QTL effects and use the QTL alleles in applied breeding (*e.g.* Bouchez et al. 2002, Ho et al. 2002, Castro et al. 2003).

To study what we perceive to be a complex biological problem it seems appropriate at this time, while we are currently building the research foundations for molecular approaches to plant breeding, to explore some of the developments and emerging concepts in the area of complexity science and its potential applications to the study of complex biological systems. Casti (1997a,b) gives an introduction to this field and some of the quantitative modeling tools that have been applied. Many of the applications of these methods have focused on the GP modeling problem. In general, but with a few notable exceptions, this work has tended to remain outside of the mainstream genomics, quantitative genetics and plant breeding communities. The current separation of these efforts may in part be explained by a combination of the recent emergence of a critical mass of research in the complexity science field, the historical momentum behind the classical quantitative genetics approaches and the recent widespread availability of high throughput technologies for the study of molecular processes in biology. With the growing availability of large data sets we anticipate and observe that there is a growing dialogue between these research fields within the genetics research community. Therefore, we are possibly positioned at a point in time where there is a broadening of the tools that will be used in the field of quantitative genetics. As one example of such an approach we will use equation (8) in combination with published QTL results to examine expected response to selection for breeding strategies. To do this we discuss the concept of the genotype-phenotype space for a trait model in terms of landscape concepts.

Fitness and Performance Landscapes

Sewall Wright (1932) introduced the idea of representing the relative performance of populations of genotypes using a landscape metaphor. He focused on evolutionary processes and therefore referred to these as fitness landscapes. From a plant breeding perspective we will generally prefer to refer to these as phenotype or performance landscapes and in some cases adaptation landscapes. In Wright's framework, high-points on the landscape represented regions of genetic space where individuals had high fitness and low-points represented regions of low fitness. The metaphor of the shape of the genotype-phenotype space as a landscape has been widely used in applications of quantitative genetics to the study of evolutionary processes, but has only been used in a more limited manner to study plant and animal breeding processes. Kauffman (1993) applied the concept of fitness landscapes to study features of the genotype-to-phenotype relationship for the *NK* model. Kauffman structured his genetic space by organizing genotypes into genetic neighborhoods based on numbers of alleles shared by the genotypes. Thus, all genotypes are arranged in what can be considered a hypercube and are one step away from all of their possible one-mutant neighbors. A fitness value is computed for each genotype and the smoothness or ruggedness of the landscape is a function of how the fitness values change with steps between genotypes on the hypercube. We recognize that the landscape metaphor does not work for everybody and may not be appropriate for all situations, but we choose to use it while remaining vigilant of its limitations.

Applying the fitness landscape framework of Kauffman, a single genotype is considered a vertex (an intersection point) within the hypercube that defines *N*-dimensional genetic space. Thus, a population of genotypes may be considered as a population of such vertices. The outcomes of the selection process may be viewed as the creation of a new population of vertices from an old population of vertices within the confines of the *N*-dimensional genetic space. This concept of genetic space may seem theoretical and abstract to the applied breeder. However, this framework can be implemented within a computer simulation environment using equation (3) to investigate any plant breeding process for a range of simple to complex GP models of traits (Podlich and Cooper 1998).

The genetic architecture of traits is a continuum: Simple to Complex Trait Genetics

A key motivating principle for organizing and intensively exploring the GP modeling problem space on a complexity continuum is to gain some theoretical insights into the differences in power that can be expected of alternative breeding strategies for a range of situations that simulate simple to complex trait genetics. Using the *E(NK)* model as defined in equation (3), levels of *E*, *N* and *K* are selected and

combined with a range of levels of heritability to stratify the complexity continuum in ways that are considered relevant to experimentally determined features of GP models. The power of different breeding strategies to achieve progress from defined starting points on the performance landscape and thus improve the trait phenotype by searching the performance landscape is compared across the levels of *E*, *N* and *K* and heritability considered.

Here we use Kauffman's landscape concept in combination with the *E(NK)* model to examine how the shape of the trait phenotype landscape changes with the genetic architecture of a trait, as determined by changes in the levels of *E*, *N* and *K*. By systematically changing the components of the *E(NK)* model we are attempting to simulate some of the context dependent properties of GP associations for traits as a continuum that ranges from simple to complex. The simple additive finite locus models are defined by the case where *E*=1 and *K*=0, thus *E(NK)* = 1(*N*:0) (Figure 1). As *E* and *K* are increased for a given level of *N*, the effects of the alternative alleles for the *N* genes become increasingly context dependent on the genotypes of other genes and on the range of environment-types in the target population of environments. Thus, context dependent effects of genes due to epistasis and gene-by-environment interaction can be simulated (Cooper and Podlich 2002).

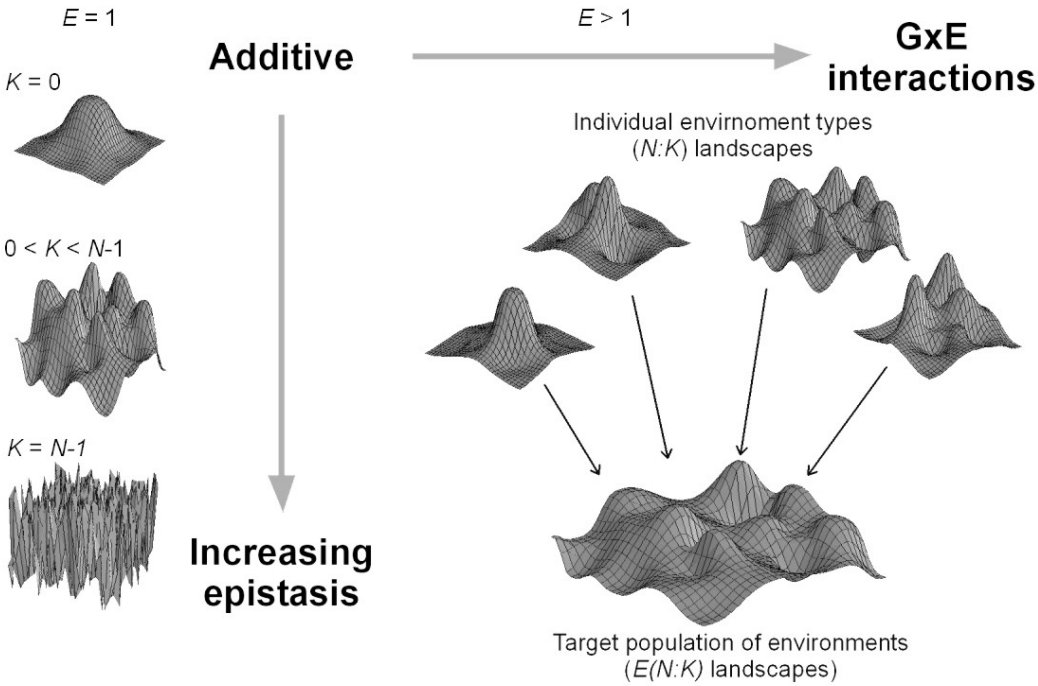

Figure 1. Schematic of performance (adaptation) landscapes for GP models simulated using the *E(NK)* model. The additive *E(NK)* = 1(*N*:0) GP model is depicted as a single peak landscape. Models with increasing levels of epistasis (i.e. from *K* = 1 to *K* = *N*-1) are depicted by an increasingly more rugged landscape surface. Models with gene-by-environment (G×E) interactions are depicted as a series of different landscape surfaces for different environment-types (*E*). The GP response surface for the Target Population of Environments is depicted as a mixture of the response surfaces from the different environment-types (*cf.* Figure 2).

Building on the landscape metaphor (Figure 1), we observe that as *E* and *K* are increased we move from a single peaked additive landscape for the *E(NK)* = 1(*N*:0) case to a multiple peaked landscape and ultimately a random landscape when *K*=*N*-1 and *E*>1 (Cooper et al. 2002). Kauffman (1993) discusses the shape of a landscape in terms of its ruggedness. We can quantify the complexity of the landscape shape by computing an autocorrelation coefficient between phenotype values for the neighboring genotypes in the genetic space for sequences of random walks across the landscape. A high value of the autocorrelation is associated with a relatively simple landscape structure and conversely a low value is associated with a complex (rugged) landscape (see Figure 2). While we are interested in the shape of the landscape for a GP model we are also interested in how effective different plant breeding strategies are at moving populations of genotypes across features of the different landscapes and in particular their power to move populations to positions of higher performance on any landscape. Thus, we construct a complexity-response plot for

any GP model by plotting the landscape autocorrelation coefficient against the response to selection for a breeding strategy (Cooper and Podlich 2002). The response can be measured in a number of ways; here we measure the change in population mean phenotype performance after a number of cycles of phenotypic (mass) selection for an ensemble of *E(NK)* models (Figure 2a). Using the complexity-response plot we observe a quantitative representation of the expectation that as the complexity of the genetic architecture of a trait increases the response to selection tends to decrease. Equally we can construct a complexity-response plot that shows the difference in the response to selection between two breeding strategies (Figure 2b). In this example marker-assisted selection (MAS) is compared to phenotypic selection (PS).

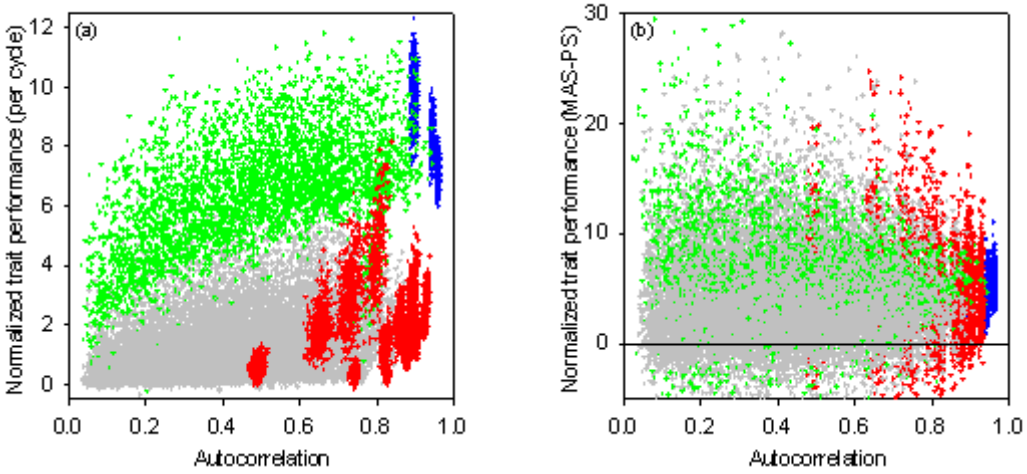

Figure 2. Complexity-response plots for an ensemble of *E(NK)* genetic models; (a) Response to phenotypic selection (change in population mean per cycle after five cycles), (b) The difference between marker-assisted selection (MAS) and phenotypic selection (PS) at cycle five (b). Each point in the ensemble represents a different genetic (GP) model implemented using equation (3); blue = additive (*E*=1, *K*=0); red=epistatic effects only (*E*=1, *K*>0); green=gene-by-environment effects (*E*>1, *K*=0); and grey=epistatic and gene-by-environment effects (*E*>1, *K*>0) (*cf.* Figure 1).

While our applications of the *E(NK)* model are defined to simulate putative properties of the ways genes interact in networks to determine trait phenotypes, it can be argued that there is nothing intrinsically biophysical about the genotype-to-phenotype mappings specified by any ensemble based parameterization of the *E(NK)* model. By drawing the effects of the gene combinations from some underlying distribution of effects we have defined genotype-to-phenotype mappings without any need to incorporate specific biological phenomena. Wherever it is feasible we seek to replace the artificial ensemble of *E(NK)* models of GP relationships with experimentally determined and validated GP models for traits. Given the diversity of the GP problem space such models will take many forms. At present some of the options available include: (1) direct parameterization or replacement of the *E(NK)* model using the results of QTL mapping studies for specific traits, *i.e.* equation (7); (2) molecular network models (*e.g.* Peccoud et al. 2004); (3) integration of separate trait mapping studies using appropriate ecophysiological models and crop growth and development framework (*e.g.* Cooper et al. 2002; Chapman et al. 2003; Reymond et al. 2003). Here we use equation (8) to superimpose some preliminary published QTL models for traits on the problem space depicted by the complexity-response plot given in Figure 2.

Figure 3 shows the results of a simulation experiment where the complexity-response plot was constructed for two traits using the results reported from QTL mapping studies (Figure 3a: Lodging resistance in wheat; Keller et al. 1999, and Figure 3b: Head blight in barley; Zhu et al. 1999). In both examples, the *explained* genetic component was parameterized by the QTL information reported in the mapping study, and the *unexplained* genetic component was parameterized by an ensemble of *E(NK)* model effects following the example given in equation (7). For the traits considered here, the number of QTL in the *explained* component of the model was 10 additive QTL (Figure 3a; Lodging resistance) and 6 additive QTL (Figure 3b; Head blight). Based on the results of the published QTL mapping studies, it was assumed that these QTL *explained* 72.6% and 14.0% of the genetic variation for the two traits, respectively, and the remainder of the genetic variation was generated by the *unexplained E(NK)* model component of equation (8). In this experiment, 50 different *E(NK)* model scenarios were considered for

the *unexplained* genetic component giving rise to 50 different parameterizations of equation (8) for each trait. The *E(NK)* models considered were a factorial combination of *E*=1, 2, 5, 10; *N*=2, 5, 12, 24, 36; and *K*=0, 1, 2, 5. Thus, the *unexplained* component of the genetic model ranged from simple (*E*=1, *K*=0; *i.e.* additive effects only) to complex (*E*=1, *K*>0; *i.e.* epistatic effects; *E*>1, *K*=0; *i.e.* gene-by-environment interaction effects; *E*>1, *K*>0; *i.e.* gene-by-environment interaction and epistatic effects). For each of the 50 parameterizations of trait performance, a complexity-response plot was constructed by considering trait response after five cycles of directional selection and computing the autocorrelation coefficient for the generated performance landscape, as described above. The performance landscape was defined by the combined effects of the *explained* and *unexplained* genetic components of equation (8). Grey circles in Figure 3 represent the complexity-response values for each of the 50 parameterizations of trait performance.

The spread of circles on each of the figure panels (Figure 3) illustrates the potential variation in the outcomes for different genetic models, in terms of the structure of the performance landscape and the expected response to selection. For the lodging resistance trait (Figure 3a), a large percentage of the genetic variation is *explained* by (known) additive QTL effects. Hence, there is a relatively small spread in the positions of the circles and the circles are confined to the upper-right portion of the complexity-response plot, indicating high levels of response and a relatively smooth performance landscape. In contrast, for the head blight trait, a relatively small percentage of the genetic variation is *explained* by (known) additive QTL effects and hence trait performance is dominated by the "*unexplained*" *E(NK)* model component of the genetic model. This results in a large spread in the positions of the circles (Figure 3b). Furthermore, there is a large difference in the complexity-response values depending on whether the *unexplained* genetic component is defined as simple (*i.e.* additive; solid square) or complex (*i.e.* epistasis, gene-by-environment interaction; solid triangle). Based on the results of this experiment, we would have significantly less confidence in predicting the outcomes from selection for the head blight trait compared to the lodging resistance trait, assuming the results of the mapping studies are representative of the genetics of these two traits.

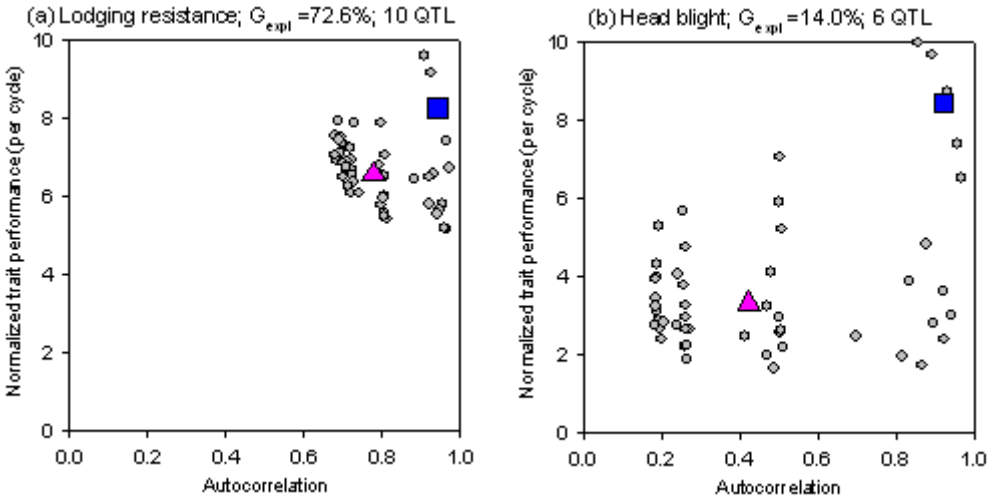

Figure 3. The complexity-response plot for two traits, where the "*unexplained*" genetic component is defined by a range of *E(NK)* models. The circles indicate the results from individual *E(NK)* model parameterizations. The solid squares represent the average results from the *E(NK)* model parameterizations that contained additive effects only. The solid triangles represent the average results from the *E(NK)* model parameterizations that contained contain epistasis, gene-by-environment interactions, or a combination of both (*cf.* Figure 2a).

The approach described above can be applied to the results of any QTL mapping study. Figure 4 shows the results of a simulation experiment where the above approach was applied to 130 different mapping studies reported in the plant breeding literature. The large variation in the outcomes emphasizes that the genetic architecture of traits can be considered a continuum from simple to complex. As illustrated by the clustering of additive models in the upper-right portion of the figure, assumptions about additivity can result in an optimistic view of the potential responses to selection for complex traits. However, a lower and perhaps more realistic representation of the expected responses to selection for complex traits is observed when different forms of context dependency (*i.e.* epistasis, gene-by-environment interaction) are

introduced into the genetic model (Figure 4). In some cases the assumptions of additivity or specific forms of epistasis and gene-by-environment interactions may be justified (*e.g.* Bouchez et al. 2002, Castro et al. 2003). In these cases we would expect to see a closer agreement between predicted and realized response to selection. In other cases the assumptions will not be justified and the agreement will be poor (*e.g.* Bouchez et al. 2002).

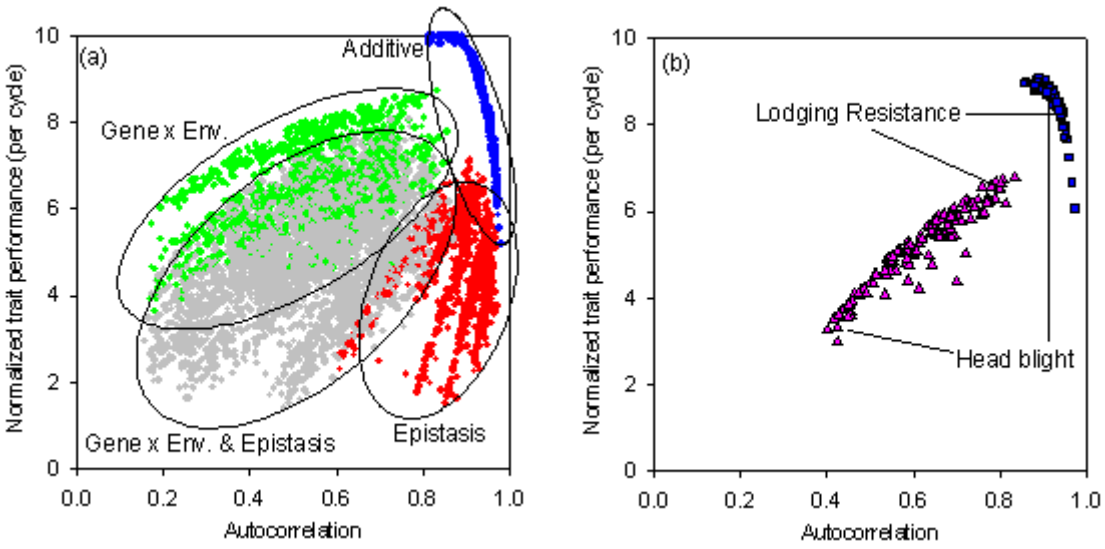

Figure 4. The complexity-response plot based on the results of 130 mapping studies, where the "*unexplained*" genetic component is defined by a range of *E(NK)* models (*cf.* Fig. 2). The results from individual *E(NK)* parameterizations are shown in (a). The average results from the *E(NK)* model parameterizations that contain additive effects only (squares) or epistasis, gene-by-environment interactions or a combination of both (triangles) are shown in (b).

Studying short-term and long-term response to selection

The majority of the selection prediction equations used in plant breeding were developed to predict expected changes in mean trait performance of a trait(s) for one cycle of selection. The equations have some applicability over a number of cycles of selection for the case where the assumption that all genes have independent and cumulative effects is appropriate. However, in the presence of epistasis and gene-by-environment interactions it is difficult to construct appropriate prediction equations for a single cycle of selection. Attempting to predict across multiple cycles of selection becomes extremely problematic.

Our interests are broad. We want to understand the predictive power that can be achieved for a breeding strategy for the continuum of simple to complex traits in both the short-term and the long-term. Using the *E(NK)* model and the components described above, we have simulated genetic and phenotypic changes for simple to complex traits for a range of plant breeding strategies. Figure 5 shows the response to selection of two of the *E(NK)* model parameterizations considered for the trait head blight (Figure 3b). The first of these genetic models is considered relatively simple (*i.e.* E=1, K=0; additive effects only) and the second is considered relatively complex (*i.e.* K>0; epistatic effects). For the scenario where only additive effects are defined (blue lines), there is a rapid increase in trait performance and accumulation of favorable alleles over the cycles of selection. Furthermore, there is little variation among the 10 independent runs of the breeding program, indicating similar trajectories have been taken in genetic space over the cycles of selection. For the model with epistatic effects (red lines), the phenotypic and genotypic response profiles are typical of what we observe for multiple peaked performance landscapes. Here, the response to selection is much slower than for the additive model and the genetic structure of the population displays greater variation from run to run. Thus, despite relatively similar values in trait performance, the populations from independent runs exhibit large genetic differences, both in the short-term and long-term responses to selection.

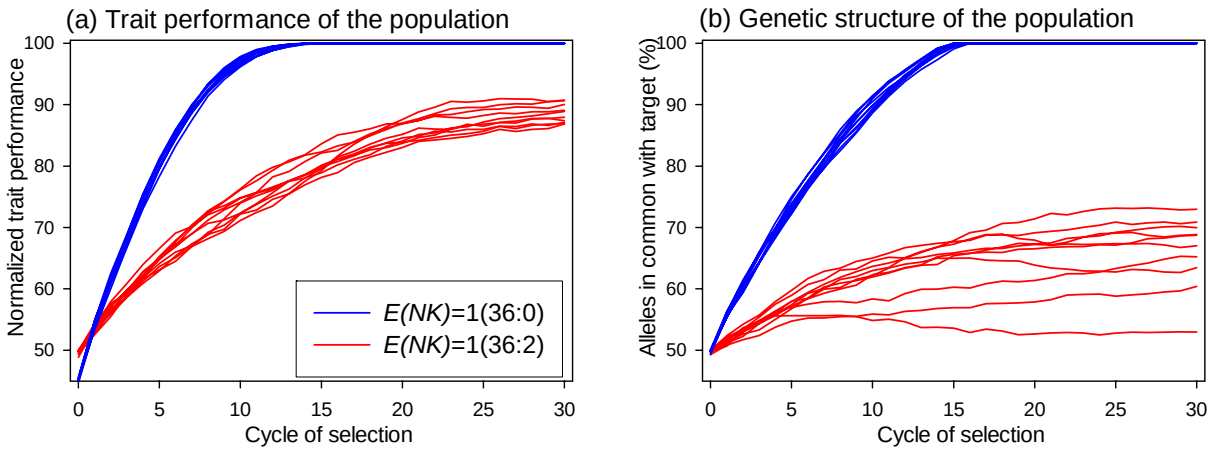

Figure 5. Response to selection for two genetic models. Each line represents an independent run of a breeding program from exactly the same starting reference population of germplasm.

Discussion

The nature of the GP relationship is fundamental to the outcomes of both conventional and molecular breeding strategies. Many have emphasized some of the complexities of the GP relationship when studying the regulation of gene expression, post-transcriptional modifications, protein structure and function, the inter-relationships among the products of gene expression within biochemical pathways, the organization and localization of pathways within cells, tissues, organs and the interplay of traits in determining organism adaptation to environmental conditions and the fitness of organisms within the context of populations of individuals. There is little doubt that viewed from a linear perspective the GP relationship is complex. We have discussed the GP model building process as a complement to our current gene discovery and functional analysis methods. There is nothing intrinsic about the experimental and quantitative gene discovery methods we use that make them the appropriate quantitative methods for building appropriate GP models. Thus, the experimental paths from organism and their traits to the gene may be different from the paths for predicting from gene to the traits of organisms. We seek a framework that enables us to begin to integrate our pieces of knowledge of the genotype-environment system and thus a strong interplay between our classical reductionist approaches to genetics and our ambitions to make strong predictions from this knowledge base to phenotypes within a complex biological system. Currently, for most of the traits of interest to us in the agricultural crops, at best we can say we have a partial picture of some of the important genes or genome regions (*e.g.* QTL). In the majority of these cases we have a list of candidate genes, some of their alternative allelic forms and a list of hypotheses about how these genes function to influence traits. Experimental testing and validation of all possible genes and hypotheses is impractical. The model-based predictions offer opportunities to prioritize and focus our experimental efforts, as has done in a number of other complex scientific and business settings.

Today most scientists are aware of the growing recognition of the importance and implications of networks for the study of biological systems. We seek to understand the simple to complex continuum of GP relationships for traits and whole organisms within a genotype-environment system context. For some this goal seems unrealistic. The motivation for investigation of network models underlying GP relationships for traits, in contrast to the additive finite locus or infinitesimal models, is quite simply the growing body of data demonstrating the networks within the cell and higher levels of organization within the organism and the genotype-environment system. Along the simple to complex continuum the additive independent locus genetic model is viewed as a reference point from which we can study the properties of network models. An important question deals with understanding the organization of these networks, do they have properties of "*scale free*" networks, i.e. many genes with $K < 3$ or 4 and a few genes with $K > 4$, or are they organized in a highly modular or structured manner with "*canalization*" versus highly unstructured random networks? We might expect emergent properties of *structured* networks could have important consequences for breeding strategies.

While we emphasize and discuss the challenges of building reliable GP models it is important to recognize that in many situations our conventional plant breeding strategies, based on selecting on the target trait phenotype, have made genetic progress for many of the complex traits in our major crops. Therefore, progress from selection can be made in the short-term and the long-term. Thus, the complexity

we face is not such that progress is not possible. The key question is whether with greater GP knowledge we can better understand, manage and direct the shape of the progress we will make in the future. Thus, a key consideration in this research field is judging the success of the GP knowledge and associated models as a basis for improving on our current capabilities to achieve response to selection for complex traits. We re-emphasize the point we made in the introduction; Ultimately it will not be sufficient to demonstrate that we can predict phenotypic variation and the phenotypic changes that result from selection using genetic information, but that this knowledge allows us to improve on the outcomes that are currently being achieved by conventional selection on phenotype alone.

References

Bouchez A, Hospital F, Causse M, Gallais A, Charcosset A (2002). Marker-assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162, 1945-1959.

Casti JL (1997a). Reality Rules: I. Picturing the World in Mathematics, The Fundamentals. John Wiley & Sons Inc., New York.

Casti JL (1997b). Reality Rules: II. Picturing the World in Mathematics, The Frontier. John Wiley & Sons Inc., New York.

Castro AJ, Chen X, Corey A, Filichkina T, Hayes PM, Mundt C, Richardson K, Sandoval-Islas S, Vivar H (2003). Pyramiding and validation of quantitative trait locus (QTL) alleles determining resistance to barley stripe rust: Effects on adult plant resistance. Crop Science 43, 2234-2239.

Cooper M and Podlich DW (2002). The *E(NK)* model: Extending the NK model to incorporate gene-by-environment interactions and epistasis for diploid genomes. Complexity 7, 31-47.

Cooper M, Podlich DW, Micallef KP, Smith OS, Jensen NM, Chapman SC and Kruger NL (2002). In 'Quantitative Genetics, Genomics and Plant Breeding'. (Ed. M.S. Kang), pp. 143-166. (CABI Publishing, Wallingford, UK).

Coors JG (1999). In 'The Genetics and Exploitation of Heterosis in Crops'. (Eds J.G. Coors and S. Pandey) pp. 225-245. (ASA-CSSA-SSSA, Madison, Wisconsin, USA).

Falconer DS (1960). Introduction to Quantitative Genetics. Oliver and Boyd.

Falconer DS and Mackay TFC (1996). Introduction to Quantitative Genetics, Fourth Edition. Longman, Burnt Mill, Harlow, Essex, England.

Fraser A and Burnell D (1970). Computer Models in Genetics. McGraw Hill Book Company, New York.

Ho JC, McCouch SR, Smith ME (2002). Improvement of hybrid yield by advanced backcross QTL analysis in elite maize. Theoretical and Applied Genetics 105, 440-448.

Kauffman, SA (1993). The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York.

Keller M, Karutz C, Schmid JE, Stamp P, Winzeler M, Keller B, Messmer MM (1999). Quantitative trait loci for lodging resistance in a segregating wheat × spelt population. Theoretical and Applied Genetics 98, 1171-1182.

Lynch M and Walsh B (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc. Sunderland Massachusetts, USA.

Peccoud J, Vander Velden K, Podlich DW, Winkler C, Arthur L and Cooper M (2004). The selective values of alleles in a molecular network model are context dependent. Genetics 166, 1715-1725.

Podlich DW and Cooper M (1998). QU-GENE: A platform for quantitative analysis of genetic models. Bioinformatics 14, 632-653.

Wright S (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Proceedings of the Sixth International Congress of Genetics. Ithaca, New York, pp. 356-366.

Zhu H, Glichrist L, Hayes P, Kleinhofs A, Kudrna D, Liu Z, Prom L, Steffenson B, Toojinda T, Vivar H (1999). Does function follow form? Principal QTLs for *Fusarium* head blight (FHB) resistance are coincident with QTLs for inflorescence traits and plant height in a doubled-haploid population of barley. Theoretical and Applied Genetics 99, 1221-1232.