Exploiting non-additive effects in genomic selection: a quadratic quantitative genetics approach

C. Chevalet[*], B. Servin[*], M. SanCristobal[*]

Introduction

The advent of Genomic Selection (Meuwissen *et al.* (2001)) has promoted a renewal of the basic approach of Quantitative Genetics, giving actual value to the notion of gene effects which had remained an abstraction until large scale genotyping became available during the last five years. Developments remain however mainly restricted to an additive statistical framework, allowing improved estimates of genetic values to be obtained.

Coping with interactions between gene effects (epistasis) seems however necessary, given the biological evidence that genes do not act independently. Gene regulatory networks, as well as metabolic networks clearly suggest that epistatic interactions should take a significant role in the statistical setting of genomic selection (Gjuvsland et al (2007); Fiévet *et al.* (2010)). Valuable approaches have been proposed to identify sets of loci that exhibit non additive effects on a trait or on several traits. Some current trends are illustrated in the following.

Population genetics theory has incorporated epistasis for a long time, but did not always suggest it should contribute much genetic variation (Hill *et al.* (2008)). Even if the additive genomic prediction of phenotypes has now proved its efficiency, it seems useful to investigate extensions of the model that could account for possible interactions between genes (epistasis) and for possible constraints on the realistic values of traits and relationships between traits (homeostasis). To do so, we propose a quadratic extension of the usual model and raise the question in a general multitrait context. The model could be the basis for a practical extension of genomic prediction, but such a general quadratic setting also suggests how a continuous selection pressure (natural selection) might lead to some co-adaptation between genetic variability and the admissible relationships between genotypes and phenotypes in which epistatic interactions play a significant role.

Searching for epistatic interactions

Many works have been carried out during the last years, making it difficult to propose a comprehensive overview of this emerging field. Only few examples of current works will be evoked here. Depending on the objective and on the experimental design, methodological efforts have been put on the development of tools aimed at detecting pairs of loci, or subsets of independent or linked Single Nucleotide Polymorphisms (SNPs), whose interactions

[*] UMR444 Laboratoire de génétique cellulaire, INRA Toulouse, BP52627, F-31326 Castanet Tolosan cedex, France.

explain the phenotypic variation of complex characters. All works were faced to the statistical questions raised by the number of combinations to explore and by model selection.

In human genetics, focus has been mainly devoted to the use of case-control designs concerning susceptibility to complex diseases. A review of methods was given by Heidema *et al.* (2006). Special attention was given to the detection of high order interactions. An approach consists in grouping loci, either single SNPs or groups of tightly linked markers (haplotypes), into subsets ("epistatic modules") so that these subsets would act additively on the trait while interactions would concern loci within each subset (Tang *et al.*, (2009)). Such approaches may be developed for *ab initio* search of interacting loci, or to test candidate groups of genes identified in a previous analysis or from biological considerations (Zhang and Liu (2007); Lee *et al.* (2007); Wan *et al.* (2009)). Selection of models remains a concern in all approaches, and some authors proposed implementing various scoring methods rather than proposing their own "best" method (Mechanic *et al.* (2008)).

In mice and in species of agronomic interest, most works were devoted to simple experimental designs (backcross, F2, recombinant inbred lines, crosses between divergent lines) for which the standard diallelic genetic model of epistasis (Kempthorne (1954); Cockerham (1954); Mao and Da (2005); Alvarez-Castro *et al.* (2008)) can be written down. In general, detection of significant epistatic interactions followed the preliminary detection of loci (QTLs) showing significant additive contributions to trait variation. The first selection of loci is then used either to reduce the search for interacting pairs, or to fix a larger but limited number of loci among which interactions are suspected. Model selection were based either on likelihood ratios (Jannink and Jansen (2001)), permutations (Yang *et al.* (2007)), various versions of Bayes Information Criterion (Yi *et al.* (2005); Yi *et al.* (2007); Bogdan *et al.* (2008); Zhang *et al.* (2008)) or penalized maximum likelihood (Boer *et al.* (2002); Zhang and Xu (2005); Xu (2007)). Another approach was tested, aimed at considering all the pairwise loci combinations and at searching for all the possible epistatic interactions effects. The method requires intensive computation and its results may be analyzed in a second step to identify potential interaction networks between loci (Ma *et al.* (2008)). Significant results were obtained in various cases, especially in poultry, with a multitrait approach of growth characters from an F2 design between White Leghorn and jungle fowl (Le Rouzic *et al.*(2008)) or body composition from an F2 between divergent lines selected for growth rate (Ankra-Badu *et al.* (2010)), or in pig (Noguera *et al.* (2009)).

However such works were mainly focused on the detection of significant interacting pairs of loci, rather than on the use of interaction effects to improve the prediction of phenotypes from molecular polymorphism. Estimates of global interactions between one locus and the genetic background have been investigated. Jannink (2007) found that, in a large proportion of simulated data, this could provide a significant improvement, and Jannink *et al.* (2009) have reviewed the question in the context of plant breeding. However, recent investigations in genomic selection suggested that accounting for interactions between gene effects did not provide any significant improvement over additive prediction (Legarra *et al.* (2008); Lee *et al.* (2008)).

Theoretical framework: Quadratic Quantitative Genetics

A general quadratic model. We consider a large diploid monoecious population under random mating. The genetic system is made up of a number ***L*** of loci, where genes have effects distributed in the population according to a multinormal distribution, defined among gametes by a mean vector ***a*** and a variance covariance matrix ***G*** (Lande (1976)). From the normal model of gene effects, we then turn to a general quadratic model for phenotypes. Phenotypes are considered for a number ***N*** of traits x^i (***i*** = ***1***,…,***N***), representative of all possible phenotypes measurable on an individual. We assume that any phenotype x^i is linked to gene effects ***(a,b)*** and to environmental effects e^i by some function $x^i = P^i(a,b) + e^i$, restricting here the model to an additive and independent contribution of environment. Considering phenotypes as continuous variables and assuming only small gene effects (as in the usual setting of quantitative genetics), we assume that these functions can be approximated by their second order development around mean genetic values. Assuming symmetry between sexes, the phenotype x^i of an individual with genotype $(\bar{a}+\alpha, \bar{a}+\beta)$ is written as

$$x^i = P^i(\bar{a},\bar{a}) + {}^T\alpha \frac{\partial P^i}{\partial a}(\bar{a},\bar{a}) + {}^T\beta \frac{\partial P^i}{\partial b}(\bar{a},\bar{a}) + \frac{1}{2}\left({}^T\alpha \frac{\partial^2 P^i}{\partial a^2}\alpha + 2\,{}^T\alpha \frac{\partial^2 P^i}{\partial a \partial b}\beta + {}^T\beta \frac{\partial^2 P^i}{\partial b^2}\beta \right) + \cdots + e^i.$$

This is re-written as follows, recalling that the means $\bar{P}^i = P^i(\bar{a},\bar{a})$ and the partial derivatives depend on the mean value $\bar{a}$:

$$x^i = \bar{P}^i + {}^T\alpha\, B^i + {}^T\beta\, B^i + \frac{1}{2}\left({}^T\alpha\, C^i \alpha + 2\,{}^T\alpha\, D^i \beta + {}^T\beta\, C^i \beta \right) + \varepsilon^i. \qquad (1)$$

where $\boldsymbol{B^i}$ is a vector of dimension ***L***; $\boldsymbol{C^i}$ and $\boldsymbol{D^i}$ are square matrices of dimensions ***L*** *x* ***L***; ***α***, ***β*** and ***ε*** are normal vectors with ***0*** means. In a large population, the first two moments of phenotypes are

$$\begin{aligned} E(x^i) &= \bar{P}^i + trace(GC^i) \\ Cov(x^i, x^j) &= 2\,{}^T B^i G B^j + trace(C^i G C^j G + D^i G D^j G) + Cov(e^i, e^j). \end{aligned} \qquad (2)$$

where $\boldsymbol{{}^T B^i}$ stands for the transpose of $\boldsymbol{B^i}$ and ***G*** is the variance covariance matrix of gene effects. Further, expressions of covariances between traits measured in relatives may be derived using the coefficients of inbreeding and of relationship. Comparing with the standard settings, terms involving the diagonal elements of matrices ***D*** correspond to dominance interaction, while other terms in ***D*** and terms in ***C*** correspond to additive x additive epistatic interactions.

Dynamics under selection. Let us assume individuals are submitted to some continuous selection process acting between the zygotic state and the adult stage. We assume optimizing Gaussian selection characterized by an optimum value **λ**, a vector of dimension the number of traits ***N***, and a positive semi-definite matrix $\boldsymbol{S^-}$ of dimensions ***N*** *x* ***N***, so that the fitness of a zygote with phenotypic values ***x*** is proportional to

$$\exp\left(-\frac{1}{2}\,{}^T(x-\lambda)\, S^-\, (x-\lambda) \right).$$

Following relationships for Gaussian genetic models (Lande (1976,1980); Karlin (1979); Chevalet (1994)), the dynamical equations relating genetic parameters in one generation to

those in the next one can be derived, neglecting third order terms. Assuming that there exists a non trivial equilibrium in which genetic variation is not exhausted (as would be in general the case for a purely additive model with no mutation), a necessary and sufficient condition for equilibrium can be shown to be: $\boldsymbol{B}\ \boldsymbol{S}^{-} = \boldsymbol{0}$, where $\boldsymbol{B}$ is the $\boldsymbol{L}\ \boldsymbol{x}\ \boldsymbol{N}$ matrix made up of vectors $\boldsymbol{B}^{i}$. This is a significant condition because $\boldsymbol{B}$ depends on the mean genetic value of $\boldsymbol{a}$ and on the phenotypic functions, while $\boldsymbol{S}^{-}$ is determined by the environmental conditions. This leads to the definition of two sets of characters, “fitness traits” and "other traits", which are derived from the original phenotypes by a linear transformation. The transformation is defined so that the selection matrix $\boldsymbol{S}^{-}$ applied to the new traits can be partitioned as

$$S^{-} = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma^{-1} \end{pmatrix}.$$

where $\boldsymbol{\Sigma}^{-1}$ is positive definite, and applies only to the fitness traits. Hence, the condition states that there is no first order (additive) effects of gene variations on the fitness, *i.e.* the additive genetic variances of fitness traits are zero. The genetic variability of fitness traits is all included in the quadratic terms; the genetic variance components for these traits are only due to the quadratic terms in $\boldsymbol{C}$ and $\boldsymbol{D}$ (Equation 2), giving rise to non zero genetic (non additive) covariances between relatives for these traits, as well as covariances between fitness traits and other traits.

Qualitative expectations. Although obtained under restrictive assumptions, the previous result suggests how the relationship between genotype and phenotype may be adapted to selective forces owing to the maintenance of non additive gene effects on traits linked to fitness. The equilibrium conditions define a subspace of the genetic space in which the links between genotypes and phenotypes are adjusted to the selection pressure of environment. In this metaphor of equilibrium between environment and genetic diversity, any departure from the equilibrium would promote changes in the genotype-phenotype interactions, develop new additive genetic effects in fitness traits and increase the genetic load. These consequences can be calculated from the model provided the main fitness traits are identified and the parameters ($\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ matrices) are known. In particular, the additive link between genotypes and phenotypes is modified under such a departure from equilibrium, since the mean genetic values, hence the values of the $\boldsymbol{B}$ matrix, are changed. Such changes are expected to decrease the efficiency of genomic prediction of phenotypes, a situation that could result from the use of genomic selection.

Checking the evolutionary relevance of the previous qualitative results would need large scale surveys with phenotypic data on many traits in many individuals. More generally, the quadratic model (1) can be used to build a genomic prediction of phenotypes that would include interaction terms. We discuss in the following possible approaches based on genome wide genotyping and genomic predictions from population data rather than from data collected in structured families.

Genomic prediction

The genomic approach of model (1) raises the usual statistical problem found when the number of unknown parameters is far larger than the number of data, since we have to deal with many traits ($\boldsymbol{N}$), many loci ($\boldsymbol{L}$) and very many interaction terms. When restricted to

additive terms (matrices $\boldsymbol{B}^i$), the model is the usual model for genomic selection for which practical algorithms are available (Lee *et al.* (2008)). The present introduction of a single "gene effect" $\boldsymbol{a}$, tranformed to an effect on traits through the matrix $\boldsymbol{B}$ does not change anything for the additive part. Additional quadratic terms, although potentially very numerous, should account for both interaction terms and for pleiotropic effects. We propose a combined approach to get computational feasibility.

Haplotype identification. A preliminary analysis of the genetic diversity of the population is to be carried out to identify shared haplotypes, and replace genome-wide genotypes (from large scale SNP chip data, for example) with a set of loci, each one being characterized by a list of haplotypes (Zhang *et al.* (2006); Sun *et al.* (2007)). Let $\mathrm{h}'_\ell(\mathrm{u})$ and $\mathrm{h}''_\ell(\mathrm{u})$ be the haplotypes carried by individual $\boldsymbol{u}$ at locus $\boldsymbol{l}$. At this locus, we assume there are several forms of the haplotype (looked at as alleles of the locus) to which "gene effects" will be assigned after genomic prediction is performed on phenotypic data.

Additive genomic prediction. Based on this reduced genetic characterization of diversity, an additive genomic prediction is performed. This needs extending available methods, such as described by Lee *et al.* (2008), to the multitrait setting. For trait $\boldsymbol{i}$ in individual $\boldsymbol{u}$, we write the additive model as:

$$x^i(u) = \overline{P}^i + \sum_\ell B^i_\ell \left(\alpha_l(\mathrm{h}'_\ell(\mathrm{u})) + \alpha_l(\mathrm{h}''_\ell(\mathrm{u}))\right) + \varepsilon^i(u) \qquad (3)$$

where $\boldsymbol{\alpha}$**(h)** stands for the gene effect associated to haplotype **h** at the corresponding locus. In this multitrait model gene effects are defined once, while effects on traits are obtained through a single parameter B^i_ℓ per locus and per trait. Without loss of generality, non zero values of B^i_ℓ may be scaled so that the variance of $\boldsymbol{\alpha_l}$ effects is set to $\boldsymbol{1}$ (*i.e.* the matrix $\boldsymbol{G}$ becomes a correlation matrix). This is expected to allow "gene effects" to be defined for each haplotype irrespective of any trait, and to provide estimates of additive genetic merits of individuals for all the traits.

Principal Components Analysis. Reducing the size of the model may be obtained through a Principal Component Analysis (PCA), as suggested by Le Rouzic *et al.* (2008). We propose to perform the PCA on the predicted additive merits of inviduals. From the previous analysis, each individual is assigned one predicted merit for each trait (even for traits that would not be available for this individual):

$$M^i(u) = \hat{\overline{P}}^i + \sum_\ell \hat{B}^i_\ell \left(\hat{\alpha}_l(\mathrm{h}'_\ell(\mathrm{u})) + \hat{\alpha}_l(\mathrm{h}''_\ell(\mathrm{u}))\right), \qquad (4)$$

where hats indicate estimated values. A PCA is performed on the $\boldsymbol{M}$'s, to get a number $\boldsymbol{N'}$ ($\boldsymbol{N'} < \boldsymbol{N}$) of principal components. Then, if the $\boldsymbol{j}$-th component $\boldsymbol{R^j}$ is written as $R^j = \sum_i \gamma_{ji} M^i$, we define "principal phenotypes" $\boldsymbol{y^j}$ for each individual ($\boldsymbol{u}$) as $y^j(u) = \sum_i \gamma_{ji} x^i(u)$ where $\boldsymbol{\gamma_{ji}}$ are the elements of the orthonormal matrix transforming the variance-covariance matrix between $\boldsymbol{M}$'s into a diagonal matrix. Conversely, original

phenotypes can be approximately recovered from the largest ***N'*** principal components: $\tilde{x}^i(u) = \sum_{j=1}^{j=N'} \gamma_{ij} y^j(u)$.

From this PCA the first components exhibit the largest additive genetic variances, as explained by genomic information, while the last ones, associated with the smallest eigen additive variances, would represent "fitness" components in the framework of the previous genetic model in which "all" traits would be considered.

Quadratic genomic prediction of principal phenotypes. The last step is identification of the parameters in the complete quadratic model, set for the ***N'*** "principal phenotypes" y^j identified from the PCA. From Equation (4), the additive genomic predictions of y's are:

$$\hat{y}^j(u) = \sum_i \gamma_{ji} \left[\hat{\bar{P}}^i + \sum_\ell \hat{B}^i_\ell \left(\hat{\alpha}_l(\mathrm{h}'_\ell(\mathrm{u})) + \hat{\alpha}_l(\mathrm{h}''_\ell(\mathrm{u})) \right) \right]$$

$$= \sum_i \gamma_{ji} \hat{\bar{P}}^i + \sum_l \left(\sum_i \gamma_{ji} \hat{B}^i_\ell \right) \left(\hat{\alpha}_l(\mathrm{h}'_\ell(\mathrm{u})) + \hat{\alpha}_l(\mathrm{h}''_\ell(\mathrm{u})) \right)$$

which may be written shortly as:

$$\hat{y}^j(u) = \bar{\hat{y}}^j + \sum_l \hat{Y}^j_l \left(\alpha_l(u) + \beta_l(u) \right)$$

where $\hat{Y}^j_l$'s take the same role as the elements of ***B***'s in the original setting (1), and where ***α***'s and ***β***'s are the gene effects assigned to the haplotypes of individual ***u***. Searching for interaction effects is then performed setting a quadratic model to explain the differences:

$$y^j(u) - \hat{y}^j(u) = \frac{1}{2} \left[\sum_l \sum_m [(\alpha_l(u)\alpha_m(u) + \beta_l(u)\beta_m(u)) C^j_{lm} + 2\alpha_l(u)\beta_l(u) D^j_{lm}] \right] + \varepsilon^j(u), \qquad (5)$$

keeping the same notations (***C***, ***D***) for the interaction terms to be estimated.

Discussion

Introducing haplotypes and assigning quantitative values to them is proposed as a way to extend current approaches to epistasis that have been mainly developed for diallelic models. It remains however to check if model (1) and its genomic implementation (Equations (3) and (5)) allows one to define gene effects that would be independent of the traits considered and that would fit experimental data. A further difficulty may arise to assign gene effects to haplotypes that are not available in the sample used for estimation.

Estimating the genetic parameters depends on the computational feasibility of the proposed approach. Following previous works aimed at identifying epistatic effects, it seems necessary to limit the search to a restricted number of loci. One way is to select loci for their likely involvement in direct (additive) effects, although epistatic interactions have been reported between loci with no individual significant effect (Yi *et al.* (2005)). Working with several traits could however be a way to include all the loci that contribute to phenotypic variation. One could follow a penalized likelihood method to select loci with a direct effect on at least one of the traits, and then search for interacting pairs (Boer *et al.* (2002); Yi *et al.* (2007)). Another approach could rely on an independent search aimed at detecting pairs of loci

showing some evidence of interaction, from a systematic pairwise search (Ma *et al.* (2008)). This independent search could be combined with an additive genomic prediction pointing to significant loci (Lee *et al.* (2008)) and – after the PCA is performed - providing the residuals to be analyzed with the quadratic model (Equation 5).

Proposing a preliminary Principal Components Analysis may be only an option. In addition to reducing the number of traits to work on, performing the PCA on additive genetic effects should result in uncorrelated residuals, so that detecting interaction terms could be more effective. However the risk is interpreting any deviation from additivity as evidence of epistasis. Some traits with high heritability are known to be poorly explained by genomic prediction, such as human height (Aulchenko *et al.* (2009)). It would be interesting to check if, in such cases, the quadratic prediction turns out to provide a better fit.

The proposed model introduces general second order interactions between gene effects, but assumes a single independent environment component. Extensions dealing with environmental effects are possible in the same line, allowing for non additive interactions between genetics and environment (Gimelfarb (1999)).

Conclusion

We propose an extension of the usual setting of quantitative genetics, aimed at a systematic search for second order interactions between gene effects. The importance of epistasis is well recognized, but its potential use in genomic selection remains to be assessed. From theoretical considerations (which will be detailed elsewhere), it seems that non additive effects should play a significant role in the phenotypic variation of traits linked to fitness. The proposed approach needs further works to check its numerical feasibility and its ability to fit actual data.

References

Alvarez-Castro, J.M. Le Rouzic, A., and Carlsborg, O. (2008) *PLoS Genetics*, 4:e1000062.

Ankra-Badu, G.A., Shriner, D., Le Bihan-Duval, E., *et al*. (2010) *BMC Genomics*, 11:107.

Aulchenko, Y.S., Struchalin, M.V., Belonogova, N.M., *et al.* (2009) *Eur. J. Hum. Genet*., 17:1070-1075.

Boer, M.P., ter Braak, C.J.F., and Jansen, R.C. (2002) *Genetics*, 162:951-960.

Bogdan, M., Frommlet, F., Biecek, P., *et al*. (2008) *Biometrics*, 64:1162-1169.

Chevalet, C. (1994*). Genet. Sel. Evol*., 26:379-400.

Cockerham, C.C. (1954) *Genetics*, 39:859:882.

Fiévet, J.B., Dillmann, C., and de Vienne, D. (2010) *Theor. Appl. Genet.*, 120:463-473.

Gimelfarb, A. (1994) *Genetics*, 138:1339-1349.

Gjuvsland, A.B., Hayes, B.J., Omholt, S.W., *et al.* (2007) *Genetics*, 175:411-420.

Heidema, A.G., Boer, J.M., Nagelkerke, N., *et al*. (2006) *BMC Genetics,* 7:23.

Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008) *PLoS Genetics,* 4:e1000008.

Jannink, J.L. (2007) *Genetics,* 176:553-561.

Jannink, J.L., and Jansen, R. (2001) *Genetics,* 157:445-454.

Jannink, J.L., Moreau, L., Charmet, G., *et al.* (2009) *Genetica,* 136:225-236.

Karlin, S. (1979) *Theor. Pop. Biol*., 15:308-355.

Kempthorne, O. (1954) *Proc. Roy. Soc. London B,* 143:103-113.

Lande, R. (1976) *Genet. Res. Camb*., 28:221-235.

Lande, R. (1980) *Genetics,* 91:203-215.

Le Rouzic, A., Alvarez-Castro, J.M., and Carlsborg, O. (2008) *Genetics,* 179:1591-1599.

Lee, S.H., van der Werf, J.H.J., Hayes, B.J., *et al*. (2008) *PloS Genetics,* 4:e1000231.

Lee, S.Y., Chung, Y., Elston, R.C., *et al.* (2007) *Bioinformatics,* 23:2589-2595.

Legarra, A., Robert-Granié, C., Manfredi, E., *et al*. (2008) *Genetics,* 180:611-618.

Ma, L., Runesha, H.B., Dvorkin, D., *et al*. (2008) *BMC Bioinformatics,* 9:315.

Mao, Y., and Da, Y. (2005) *Genet. Sel. Evol*., 37:129-150.

Mechanic, L.E., Luke, B.T., Goodman, J.E., *et al*. (2008) *BMC Bioinformatics,* 9:146.

Meuwissen, T.H.E, Hayes, B.J., and Goddard, M.E. (2001) *Genetics,* 157:1819-1829.

Noguera, J.L., Rodriguez, C., Varona, L., *et al*. (2009) *BMC Genomics,* 10:636.

Sun, S., Greenwood, C.M., and Neal, R.M. (2007) *Genet. Epidemiol*., 31:937-948.

Tang, W., Wu, X., Jiang, R., *et al.* (2009). *PLoS Genetics,* 5:e1000464.

Wan, X., Yang, C., Yang, Q., *et al.* (2009) *BMC Bioinformatics,* 10:13.

Xu, S. (2007) *Biometrics,* 63:513-521.

Yang, J., Zhu, J., and Williams, R.W. (2007) *Bioinformatics,* 23:1527-1536.

Yi, N., Shriner, D., Banerjee, S., *et al*; (2007) *Genetics,* 176:1865-1877.

Yi, N., Yandell, B.S., Churchill, G.A., *et al*. (2005) *Genetics,* 170:1333-1344.

Zhang, M., Zhang, D., and Wells, M.T. (2008) *BMC Bioinformatics,* 9:251.

Zhang, Y., and Liu, J.S. (2007) *Nature Genetics,* 39, 1167-1173.

Zhang, Y., Niu, T., and Liu, J.S. (2006) *Am. J. Hum. Genet*., 79:313-322.

Zhang, Y.M., and Xu, S. (2005) *Heredity,* 95:96-104.