

## Beyond Genomic Selection

B.P. Kinghorn

University of New England, Armidale, Australia

**ABSTRACT:** Use of genomic information to directly select for identified QTL has essentially failed, working only for traits affected by one or a few QTL. We resort to additive statistical fits and empirical extrapolation, with no inference of mechanism let alone identification of QTL genotypes. Although this works very well for many scenarios, we could generally do better, especially for traits involving high levels of non-additive variation. However, even in these cases we can probably only work with traits of relatively few QTL. There may be some hope here, if QTL interacting strongly masquerade as many more QTL under our additive statistical fits. Biologically inspired mechanistic models may compete favourably with statistical models involving epistatic parameters, but great effort would be involved. There is an emerging need for systems to ensure ongoing phenotyping for genomic calibration, and proposals for this are discussed. The development of in-vitro reproductive technologies would greatly increase the impact of genomic information, with selection between zygotes and even between gametes providing new levels of genetic information and gain.

**Keywords:** Genomic Selection; Gene interaction; Epistasis; Biological modeling; Reproductive technologies

### Introduction

As animal breeders we are principally engaged in making genetic change. We do this through the passive route of managing animal selection and mate allocation, although there are options to take the active route involving recombinant DNA. This paper focuses on the passive route.

Using the passive route to making genetic change we must exploit existing genetic variation. We have classically used this variation as expressed in phenotype to build systems for making genetic change, but we now have genomic information to do this with potentially higher fidelity.

There are some benefits of genomic information that are not covered here, such as estimating breed composition, pedigree inference and verifying parentage. This paper concentrates on the extent to which genomic information could conceivably be used to more effectively make genetic change, and this of course must involve some conjecture.

### What could genomic information tell us?

For simplicity, let us assume that we have full sequence information for the nuclear chromosomes for all animals of interest, and this is our 'genomic information'. To help identify possible limitations in the value of genomic information, we can ask four questions, and make associated comments:

**1. To what extent does nuclear DNA dictate genetic merit of the host individual?** We ask this question because it relates to the asymptotic value of genomic information.

The phenotypes of the extranuclear organelles will impact both genetic merit and environmental deviations.

DNA methylation, histone packing and chromosome folding patterns will impact on expression of genetic merit. These things could be largely dictated by nuclear DNA, but possibly the nuclear DNA of parents and other ancestors as well as of the host itself.

If present, GxE interaction will render genetic merit dependent on the environment, for identified and unidentified environmental components.

The answer to this question is probably that the great majority of genetic variation is dictated by nuclear DNA.

**2. To what extent does nuclear DNA dictate transmissible genetic merit of the host individual?** We ask this question because it relates to the asymptotic value of genomic information under simple breeding program designs that exploit additive effects alone.

*Transmissible* can only be defined with knowledge of the target population. For progeny, some epistatic effects are included, for distant descendants under random mating epistatic effects are essentially absent.

Unlike the previous question, it is probably not possible to give a mechanistically derived answer to this question – it depends on factors other than biological mechanism.

However, the answer to the previous question can be used here for most simple breeding designs - a fraction of full genetic merit can be transmitted. In terms of variation in full genetic merit, that fraction is narrow-sense divided by broad-sense heritability,  $h^2/H^2$ .

### **3. To what extent can genomic information inform us about the genetic merit of the host individual?**

This depends critically on the trait(s) involved. Simply-inherited traits involving mutations at one or a few loci can be fully informed. However we currently perceive most production traits to involve many loci – they are complex traits.

For complex traits we could make empirical statistical fits, typically using a linear model with sequence information and a range of additive and non-additive effects on the right-hand side. But we could also make mechanistic fits, using biological models of trait expression. The former is limited in power but simpler to apply, the latter is asymptotically perfect (where the model adopted reflects life itself) but can be hopelessly challenging. The contrast between these two will be discussed later.

### **4. To what extent can genomic information inform us about the transmissible genetic merit of the host individual?**

We currently use statistical fits based on the average effects of alleles, with few exceptions, and so this question is now being answered routinely through estimates of accuracy of genomic estimates of breeding value (gEBVs). Answers to the first three questions would help us to understand the potential value of deviating from this course, accommodating dominance and epistatic effects in our statistical fits, or indeed exploiting some form of mechanistic fit.

#### **Estimating and exploiting non-additive genetic effects**

If we find that the answer to question 4 is close to 100% accuracy of gEBVs, for traits of interest, then the end of the road has not necessarily been met. This is only the correlation of the criterion to additive genetic merit.

Consider traits showing strong non-additive variance, such as fertility traits in dairy cattle. Palucci et al. (2007) estimate variance components that lead to  $h^2$  values of {0.1, 0.005, 0.011 and 0.067} and  $H^2$  values of {0.74, 0.049, 0.04 and 0.343} for traits {Age at first calving, Heifer non return rate, cow non return rate and interval from calving to first service}. In this case  $H^2$  values average 6.5 times  $h^2$  values!

We are unlikely to fully capitalise on this extra source of merit, for reasons related to: finding the information required; and limitations in what can be achieved in breeding program design. However, it certainly seems worth investigation!

We can use genomic information to estimate non-additive effects using statistical fits. This is done for intra marker-locus dominance by Zeng et al. (2013). However the good approach adopted seems unlikely to cover epistatic effects across the genome without masses of data. The Kernel regression method of Gianola et al. (2008) can give some angle on genome-wide epistatic effects, through management of the relationship between genetic distance and

covariance for epistatic effects. However this approach seems unlikely to capture strong inter-locus epistatic effects that seem likely for many traits – life is too complex for them to be unimportant.

Can we use mechanistic fits? This essentially means “building biological models of gene action and interaction - models which have the power to predict ideal genotypes across many loci. This is quite different from the current wide use of biometrical models to predict genetic values - biological modelling provides a mechanistic prediction of ideal genotypes across loci, not an empirical extrapolation which is typical of current predictions of genetic values. As QTL become identified at an increasing rate, there is the basis for an escalation of information on gene action and interaction. This will help with the modelling of biochemical pathways, hormone-receptor interactions, etc. However, epistasis, homeostasis, and canalisation will make such modelling of life processes very complex indeed.” (Kinghorn, 1996).

After two decades we still seem far from such dreams. We need some understanding of how life works, and as such this is a very ambitious area. Systems Biology has grown notably, but we need to connect this to genetic variation as revealed by genomic information. Success in this area implies a somewhat finite number of QTL affecting genetic variation, and current thinking is that most traits are affected by a very large number of QTL. Accordingly, mechanistic approaches may well be restricted to the relatively few traits thought to be controlled by few QTL.

Coat colour is a simple example. For many systems we have a pretty good understanding of the biological model, and many desired outcomes can be reached by only one or a few genotypic configurations of the loci involved. The task might then be somewhat different from improvement of an additively inherited polygenic trait, as we now have an ideal genotype to target. We could aim for the top of the hill, rather than climbing the prevailing steepest slope. If the ideal genotype involves heterozygosity at one or more loci, then some structured breeding design will be required to optimally approach and maintain that genotype. Even without heterozygosity at the optimum, epistatic interactions could be such that multiple lines would be involved in developments towards fixation of ideal configuration across loci.

For traits involving a low to moderate number of QTL, phenotypic information is required ideally not just for the trait(s) of interest, but also for relevant low-level phenotypes, such as levels of substrates and gene- and protein-expressions involved in pathways towards trait expression. Models can be developed from different resources, including both experimental intervention without consideration of genetic variation, and observation of the impact of different genotypes.

#### **Can we predict the ideal genotype?**

First we might ask, how far are we from the ideal genotype? This was illustrated by Kinghorn (2011), with a

simulated genome including 10,000 QTL with additive effects  $a$  sampled from a gamma distribution and dominance effects  $d$  sampled uniformly between 0 and  $a$ , but no epistasis. With no overdominance the ideal genotype is fully homozygous, and its merit is illustrated in Figure 1. Progress over 25 generations is also shown in a program using genomic information to target total genetic merit in crossbreds. It can be seen that the ideal genotype is very much above the merit in the breeding program, and would only be approached after many hundreds of generations.

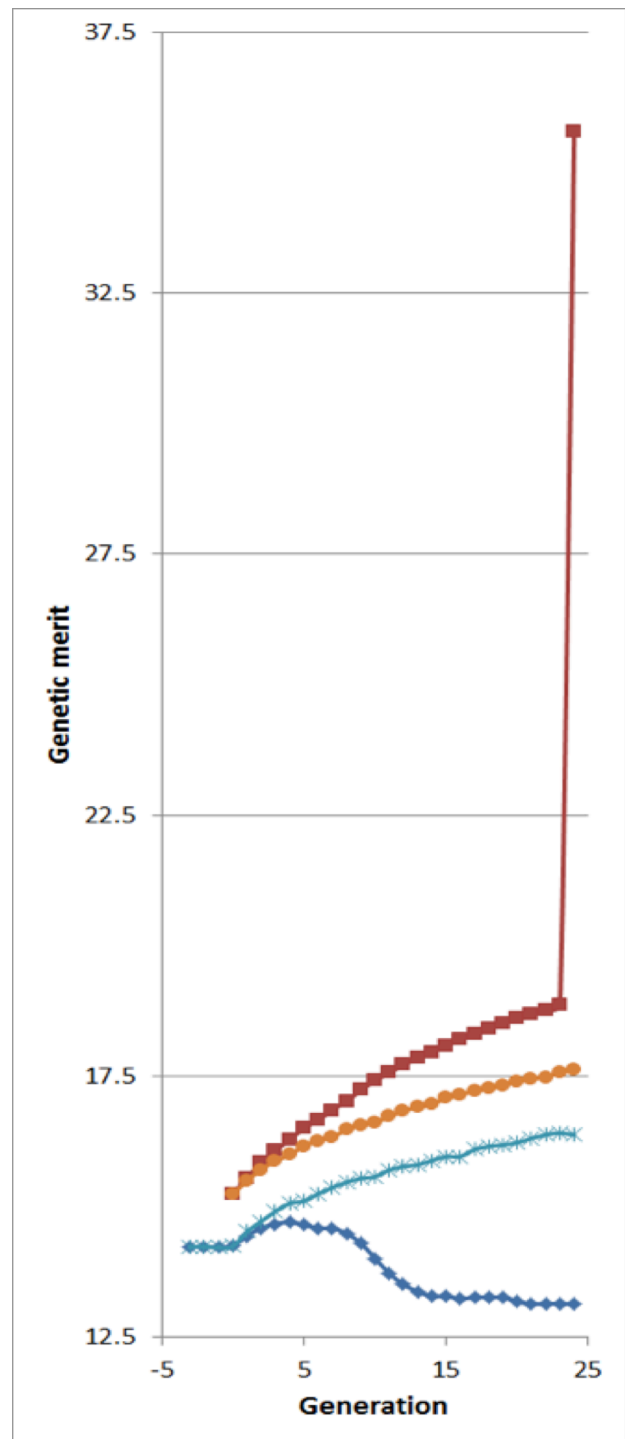
Notable increase in heterosis is evident in the RRGs line (Figure 1) despite full homozygosity for the ideal genotype. Conceptually, the steepest prevailing slope is being climbed, which exploits heterozygosity in the crossbred, rather than aiming directly at the top of the hill. Thus knowledge of the ideal genotype is here somewhat irrelevant. This illustration underlines that we need systems with relatively few QTL involved to make it worthwhile predicting and targeting the ideal genotype. Even if the model of fit were correct, the way we use it is not, unless we fully account for both short-term and long-term objectives and tactical methods to appropriately target them.

Adding even moderate involvement of between-locus interaction, it can be imagined that knowledge of the ideal genotype becomes even more irrelevant. Average effects of allele substitutions could be considerably different in the current population compared to the ideal genotype (shifting windows of genetic background – Barton and Keightley, 2002). The exception may be cases with relatively few QTL involved. However, even with so few QTL, statistical models involving eg.  $aa$ ,  $ad$ ,  $da$  and  $dd$  terms for pairs of loci seems likely to be useful only for the prevailing allele frequencies in the current populations. These are not ‘cause and effect’ models based on perceptions of reality, but somewhat abstract fits based on associations. Thus epistatic effects and variance components estimated in this way could relate poorly to biological gene interaction (Carlborg and Haley, 2004).

This is where biologically-inspired models may be more useful for predicting ideal genotypes and setting routes towards that goal: Relatively few QTL in finite systems, ideally with some prior understanding of the biological systems involved.

With no prior information there may be value in some form of network SNP-association analysis, whereby a search algorithm looks for groups of SNPs (possibly from a preselected set following more standard analysis) plus a non-linear network model that leads to a good fit to phenotypic data. This could involve, for example, sets of differential equations using SNP genotypes as part of the independent set of variables, with output being a liability score for phenotype. An evolutionary algorithm would then find both the model and the parameters by maximising the correlation between liability and observed phenotype.ww

### Ongoing calibration of genomic information



**Figure 1: Predicted merit for the best genotype (35.6 units) in relation to a breeding program using reciprocal recurrent genomic selection (RRGS) for performance in crossbreds (top line) derived from purebreds (bottom line). The second and third line from the bottom are purebreds and crossbreds under normal genomic selection, (From Kinghorn et al., 2011)ww.**

The most likely visage for genomic selection in the foreseeable future is more of the same: Presumption of very many QTL for most traits of commercial interest and consequent lack of ability to get a handle on these QTL *per se*. In this way, we use genomic information to track moderately large chunks of DNA. Within a few generations these recombine to the extent that we cannot rely on old calibra-

tions, and must maintain recording of phenotype and some form of ongoing calibration. The perpetual genomic key only works for simply inherited traits.

For closed breeding lines with  $N_e$  of say 50, we can get away with relatively few SNPs routinely genotyped in animals with phenotypes for calibration. If we were to use a QTL mapping paradigm for genomic selection we could probably get away with many fewer QTL (say 3,000 after imputation), but with many markers available, the SNP association paradigm seems to work sufficiently well.

However for larger populations and migration across breeding groups, as in dispersed national evaluation programs for pigs, beef and sheep, we need extensive and ongoing phenotyping and genotyping within breeds. This is an across-enterprise activity that was first launched largely through support from governments and industry organisations, and usually without understanding of the ongoing requirements. So there is growing recognition of need to migrate this activity out of the public-good sector. There is also a need to mitigate the reduction of phenotyping in industry that occurs with the success of genomic selection. This section considers two approaches.

**Service providers buy phenotypes and sell gEBVs.** In this scenario, service providers buy the phenotypic information of genotyped animals from breeders and producers. Pricing can be objectively set according to:

- Breed and genetic position in the breed: Higher price for less well covered breeds/lineages, and higher price for better gEBV sales prospects.
- The traits included.
- Quality of the phenotypes, e.g. as determined by deviation from their prediction based on their genotype and the prevailing genomic key.

The service provider uses the phenotypes and genotypes to make gEBVs, and then sells these gEBVs to breeders. These gEBVs would probably be exclusive to the service provider. The service provider can use the gEBVs to pay for phenotypes where possible.

**The service provider hosts an information marketplace.** In this scenario, a breeder would use a web site to order genomic tests, with the option to click checkboxes to choose the source(s) of phenotypes on high density genotyped animals to be used in the calculation of gEBVs for this breeder. For whatever choice is made, the Genomic Relationship Matrix will be the same (~breed-wide/national), and only the data vector will depend on choice, and there is a custom iteration to solve for the breeder's gEBVs. This is for a one-step approach, but specific genomic keys could be made under a Bayes approach - more computation but not a lot more complexity.

Whenever there is a change in what source-selection boxes are checked, there is an update on the cost of the phenotypic information and the predicted gEBV accuracy it will yield. This information can be displayed in

an accuracy versus cost graph with each combination of sources chosen shown as a single point. Presenting estimates of accuracy in this way may seem risky, but it is similar to predicting EBVs, as done today: breeders should know that accuracies are not perfect. Of course, the service providers method would be published for scrutiny.

The sources of information come typically from progressive breeders. They now enter a market to sell genetic information, not just genetic seedstock. This is another way to make money out of running a good breeding program. They sell their information based on figures (predicted contributions to accuracy for the prevailing customer), and also based on their reputation – the same as for selling seedstock with EBVs.

There is no basic constraint on pricing (the marketplace will determine that), but there is a pattern of pricing that the system determines. For example, for just one source of information chosen the full rate is payable, but for two or more sources chosen, there is a reduced proportion payable for each as a technical function of predicted contributions to accuracy from each source. Sellers can opt to discount their base pricing for more distantly related target stock, and the system helps them to implement that sensibly, but this would not be compulsory.

Publically funded reference populations may indeed be suppliers, but likely to wind down their contribution over time. This gives a natural and competitive basis to phase out public funding. There is no system of royalties and the likely associated suite of ploys by clients to get around the rules to increase their benefits and reduce their contributions. Subject to control of cost patterns, the system is a free market, with increasing incentive for progressive players to supply information as the overall supply diminishes (ie. a self-correcting system).

### **Genomic information to manage diversity**

Genomic information can replace pedigree information in estimating coancestry. Moderate marker densities are sufficient to outperform pedigree information (Gómez-Romano et al., 2013).

Clark et al. (2013) used genomic coancestry to show impact on the performance of optimal contributions selection in simulated breeding programs. The one scenario to benefit is where there is genomic selection among full sibs. With some emphasis on reduced coancestry, there is a tendency to co-select full sibs that are genomically less related compared to the average relationship for the family.

This is already being implemented in such scenarios using mate selection software that includes optimal contributions features. However, this approach would be of notable benefit if and when we start to make selections among large full-sib families of embryos or cell lines, and discussed below.

## The synergy between genomic selection and reproductive technologies

Much of the value in boosting fecundity through reproductive technologies comes from reduced generation intervals. The downside here is low accuracy evaluation on juvenile candidates, with phenotypic information coming typically from the parental generation.

Genomic selection can play a major role here, giving accurate evaluation on juveniles that have genotypes but no phenotypes. This does not require phenotypes on close relatives, and so we can contemplate selection among embryos from juvenile females (Georges and Massey, 1991), with the nearest phenotypes being two generations away. If we exercise extreme boosting to generate say 100 embryos per female, followed by recovery of selected individuals via nuclear transfer, due to lack of embryo viability, then selection intensity within families becomes large, and this is particularly valuable as most variation exists within families. In a simple test, the author calculates more than five-fold increase in rate of genetic gain with 100 embryos tested per juvenile female, compared to a normal genomic selection program.

We can also contemplate developing cycles of in-vitro sexual propagation, including meiosis and zygote formation, with selection among large numbers of zygotes or zygote cell lines using genomic information (Haley and Visscher, 1998; Kinghorn, 1996).

Selection among gametes or haploid cell lines using genomic information would be even more powerful (Kinghorn, 2010). This effectively accesses twice as much additive genetic variation due to covariances generated between the gametes that contribute to zygotes. Selecting the best sperm out of 10, and using this to fertilize the best egg out of 10, gives as much selection differential as selecting the best zygote out of 1000.

This approach would also help exploit non-additive effects, and the generation of targeted genotypes across many loci.

This might not only be done for the purposes of genetic improvement. We could generate suites of targeted QTL genotypes that are predicted to provide weak-link information for biological model building – to iteratively build biological models of QTL (inter)action (Kinghorn, 1996). Of course, any such developments would have to compete with recombinant DNA techniques.

### Conclusion

Genomic information promised silver bullets for simple and effective genetic improvement. However we have come to realize that life is perhaps more complex than we had hoped, and that with current ideas and methods we have to maintain a high level of effort to exploit genomic selection in our breeding programs.

With few exceptions, we exploit genomic information not through biological mechanisms revealed, but through somewhat abstract statistical fits and associated extrapolations.

However, there is still opportunity to dig deeper. Maybe, for some traits, epistatic interactions make the relatively few QTL involved masquerade as very many QTL in the additive statistical analyses that we carry out. Maybe we can discover this and start to build models that are more mechanistic in nature, and more transportable to other populations and genetic backgrounds. Such developments do not seem to be just around the corner.

The development of in-vitro reproductive technologies would greatly increase the impact of genomic information. However, we would need to have frequent realisations of in-vivo animals and their measured phenotypes to maintain genomic accuracies and stay on course.

In any event, we will need to increase and maintain effort to implement appropriate breeding program design to both exploit new opportunities, and to help maintain genetic diversity and economic viability.

### Acknowledgement

Funding from Australian Research Council Discovery Project DP130100542 is acknowledged. The author thanks Cedric Gondro, Sam Clark and Julius van der Werf for stimulating discussions.

### Literature Cited

- Barton, H.B. and Keightly, P.D. (2002). *Nature Reviews (Genetics)* January 2002. 3: 11.
- Carlborg, Ö., and Haley, C.S. (2004). *Nature Reviews (Genetics)* August 2004. 5:618-625
- Gianola, D., Fernando, R.L. and Stella, A. (2008). *Genetics* 173:1761-1776.
- Clark, S.A., Kinghorn, B.P., Hickey, J.M. et al. (2013). *Genet. Sel. Evol.* 45:44
- Kinghorn, B.P. (1996). Harald Skjervold Symposium. Agricultural University of Norway, 23 August 1996. *Acta Agric. Scand. Sect. A, Animal Sci.* 1998. 28: 27-32.
- Kinghorn, B.P. (2010). *Mating Systems*. *Encyclopedia of Animal Science*. 2nd ed. Pages 744-747.
- Kinghorn, B.P., Hickey, J.M. and van der Werf, J.H.J. (2011). Paper 1. 7th Eur. Sym. *Poultry Genet.*, 5-7 October 2011, Peebles Hydro, Scotland.
- Georges M. and Massey J.M. (1991). *Theriogenology* 35: 151-160.
- Gómez-Romano, F., Villanueva, B., Ángeles Rodríguez de Cara, M. et al. (2013). *Genet. Sel. Evol.*, 45:38
- Haley, C.S. and Visscher, P.M. (1998). *J. Dairy Sci.* 81:85-97.
- Palucci, V., Schaeffer, L.R., Miglior, F. et al. (2007). *Genet. Sel. Evol.*, 39: 181-193
- Zeng J., Toosi A., Fernando R.L., et al. (2013). *Genet. Sel. Evol.* 45:11.