The Four Horsemen Of Genomicalypse: Fuzzy notions in genomic selection

A. Legarra

1 UMR 1388, INRA, 31326 Castanet Tolosan, France
Email: andres.legarra@inra.fr (Corresponding Author)

Introduction

Here I present a non-Final Judgment on the theory supporting Genomic Selection. The number of publications concerning genomic selection totals > 2000, yet to my taste there are many unsolved questions. In particular, in spite of great efforts, current theory is unable to explain and predict accuracy from a priori parameters such as pedigrees of to-be genotyped individuals, heritability, and LD statistics. The role of close vs. far relationships is particularly poorly understood.

In order to understand how genomic selection works, a few concepts need to be clarified. Although these concepts are present in most researchers’ heads, the lack of a clear formalization leads to barriers for sharing, understanding and development. This is my brief, personal review on such four topics. I apologize for self-centred bias and omissions.

Linkage disequilibrium

The notion of linkage disequilibrium or gametic phase disequilibrium has a clear definition in the sense of “deviations of gamete multilocus frequencies from expected values”. However, the translation of this notion to rigorous mathematical definition that lead to usable metrics is notably difficult. Usual statistics () considering two loci at the same time are unusable to assess prediction ability, because for any realistic situation a set of QTL will be captured by a set of markers, in other words, the collinearity between the QTL incidence matrix \( Q \) and the marker incidence matrix \( Z \) can not be described by two-loci statistics. And the generalizations of statistics describing LD to more than two loci are generally too complex. In addition, there is basically no theory available to describe the evolution of LD, besides very general results (Sved 1971). There is in addition no theory to disentangle family LD from “far” (whatever this means) LD. The theory of junctions or the coalescent are unusable for complex pedigrees or selection. So, we have problems to describe LD, to quantify LD and to LD.

This has consequences for the understanding of genomic accuracy. The usual assumption “at least one SNP is in LD with each QTL” seems to mean little in practice. It would seem that every QTL is in (possibly very low) LD with every marker; this may explain why increasing marker density beyond 50K seems to add little to genomic accuracy. In linear model terms, a QTLBLUP would be whereas a SNPBLUP would be . SNPs are good predictors of breeding values if’ (i.e. true QTL relationships are well described by genomic relationships). This is not the same as which describes “each QTL is tagged by a marker”. The latter can be checked (conceptually) by two locus LD. However, to attack the similarity of we need multilocus measures of LD.

Relationships
VanRaden (2007, 2008) first realized that a crossproduct of marker incidence matrices $Z$ provides an estimator ($\mathbf{G}$) of realized relationships, and this crossproduct provides an elegant connection between marker-based evaluations and relationship-based evaluations. Another estimator ($\mathbf{A}$) uses linkage analysis to infer IBD relationships (Luan et al. 2012). But maybe the notion of relationships needs to be clarified. We have expected IBD relationships conditional on the pedigree ($\mathbf{A}$), real unobserved IBD relationships conditional on the pedigree and on unobserved meioses events (say ). Hill and Weir (2011) for entire genomes and simple pedigrees, and Garcia-Cortes et al. (2013) for a single locus and arbitrarily complex pedigrees, described the distribution of . For instance, in absence of inbreeding, between father and offspring there is no variation; the relationship is always 0.5. Lack of adjustment of $\mathbf{G}$ to expected values in $\mathbf{A}$ (e.g. ) is often interpreted as errors in $\mathbf{G}$, not in $\mathbf{A}$. In addition, $\mathbf{A}$ is undefined when there are several base population differing in average genetic value (i.e., unknown parent groups).

However, the notion of identical by descent relationships (Wright 1922; Malecot 1948) assumes that all animals are traced back up to an infinite unrelated population. This is of course a false assumption, first because pedigrees are finite and incomplete, second because such a quasi-infinite size population has never existed in livestock (although it does exist in species such as Drosophila or Populus. We have suggested (by statistical considerations) to consider markers as the cornerstone of relationships, the philosophy being that they are an objective measure of similarity that does not depend on recorded or unrecorded events. Then, pedigree information is used to project this genomic relationships to other animals in the pedigree (Christensen 2012; Legarra et al. 2015). This automatically makes $\mathbf{G}$ and $\mathbf{A}$ comparable. As by-products, we obtain relationships within and across base populations in the form of statistics (Garcia-Baccino et al. 2017). The theory seems solid, yet this is just a proposal.

We know in addition that GRM based on crossproducts, or on IBS, are equivalent (Strandén and Christensen 2011; Garcia-Baccino et al. 2017), which means that they describe equally well the population. So, what do we mean by relationships?

**Effective number of segments**

There is quite good an agreement on the importance of the “equivalent number of segments” ($\mathbf{A}$), which would describe the “closeness” of the information provided by the markers to a single locus model (or to an infinitesimal model ($\mathbf{A}$). This is useful, among other things, to estimate likely differences between expected ($\mathbf{A}$) and realized ($\mathbf{G}$) relationships, which in turn leads to a priori estimates of genomic accuracy (Goddard et al. 2011; Wientjes et al. 2013). Ideally, this is a population parameter which describes (another sloppy wording) the size and distribution of segments in linkage disequilibrium. The number of segments increases (segments split) and decreases (two segments can join again) with meiosis and also with drift (loss of original segments). This is the little explored theory of junctions e.g. (Stam 1980; Martin and Hospital 2011). All in all, it would seem (but to my knowledge this remains to be shown) that a, probably temporary, equilibrium can be reached, and therefore could be interpreted as a populational parameter of an existing-population at a given point in time, i.e. all Holstein cows alive in 2017, like we can define their genetic variance (Sorensen et al. 2001; Legarra 2016; Lehermeier et al. 2017). This parameter would, in turn, define for each
pair of relationships in the pedigree, in the same way as we can define the covariance as a variance times a relationship.

In my view, this is a very important point to develop the theory. Some authors (Wientjes et al. 2013) choose to call to a measure of discrepancy between A and G for particular data sets or groups of individuals. Basically, . For a population composed of several relationships, this will give different numbers of depending on the constitution of the data set, and to me this does not help. Consider an ideal genome of 30 chromosomes of 1 Morgan each. Following (Hill and Weir 2011) between father and offspring, for fullsibs, for halfsibs and for cousins. So, for the same population, depending on the constitution of the data set, we will get different definitions of. This contrasts with the generality and elegance of describing the structure of a pedigreed population by a genetic variance and two tabular rules.

One of the problems of writing a proper definition of is the lack of a theory of junctions for complex pedigrees. Even for a single locus, the variance of realized relationships is a complex function of the pedigree (Garcia-Cortes et al. 2013). This a priori variance of G-A agrees well with observed deviations (Wang et al. 2014), which may lead to populational estimates of which are better than a crude descriptor of observed G-A.

Can we do Realistic Simulations?

Simulations are useful to learn, teach and in general they show promising ways. However, they also lead to pitfalls. I will cite a few conclusions from simulations that turned out wrong (so far) in real data:

- Use of increasing marker density leads to more accurate predictions and to accurate across-breed predictions
- Bayesian regression models are better than GBLUP
- Accuracy does not depend on closely phenotyped animals
- Accuracy is stable over generations

This shows that we don’t know how to simulate realistic genomes. For many animal breeders (including myself), the mental model of the genome is a series of contiguous boxes (loci) that may or not recombine, and each box has a few alleles with definite action on the phenotype. This leaves no room for complex variants (insertions, deletions), for regulatory elements (enhancers, promoters), for fuzzy locations (e.g. coding regions + promoters + enhancers +...). It also assumes an oversimplified functional gene action (mainly additive). Although we know that the genetic variance is largely statistically additive (Hill et al. 2008), and we’re interested in statistical effects, this does not mean that QTL have functional additive effects. The existence of genomic correlations (Karoui et al. 2012; Porto-Neto et al. 2015) shows the presence of GxE but also GxG interactions across genetic architecture in different populations. Even within a population, the genetic architecture changes with time resulting in a genetic correlation across decades that can be as low as 0.72 for Milk Yield (Tsuruta et al. 2004). Although these phenomena are unpredictable and very difficult to simulate, we need not to draw too many conclusions from our simulations.

List of References

Christensen O. F., 2012 Compatibility of pedigree-based and marker-based relationship
Garcia-Baccino C. A., Legarra A., Christensen O. F., Misztal I., Pocnrc i., et al., 2017
Metafounders are related to Fst fixation indices and reduce bias in single-step genomic
Garcia-Cortes L. A., Legarra A., Chevalet C., Toro M. A., 2013 Variance and Covariance of
Actual Relationships between Relatives at One Locus. PLoS ONE 8: e57003.
Goddard M. e., Hayes B. j., Meuwissen T. h. e., 2011 Using the genomic relationship matrix
Hill W. G., Goddard M. E., Visscher P. M., 2008 Data and Theory Point to Mainly Additive
Genetic Variance for Complex Traits. PLOS Genet. 4: e1000008.
Hill W. G., Weir B. S., 2011 Variation in actual relationship as a consequence of Mendelian
sampling and linkage. Genet Res Camb: 1–18.
Karoui S., Carabaño M. J., Díaz C., Legarra A., 2012 Joint genomic evaluation of French
Legarra A., Christensen O. F., Vitezica Z. G., Aguilar I., Misztal I., 2015 Ancestral
Relationships Using Metafounders: Finite Ancestral Populations and Across
Legarra A., 2016 Comparing estimates of genetic variance across different relationship
Lehermeier C., Campos G. de los, Wimmer V., Schön C.-C., 2017 Genomic variance
Luan T., Woolliams J., Odegaard J., Dolezal M., Roman-Ponce S., et al., 2012 The importance
Evol. 44: 28.
Martin O. C., Hospital F., 2011 Distribution of Parental Genome Blocks in Recombinant
Genomic correlation: harnessing the benefit of combining two unrelated populations
Sorensen D., Fernando R., Gianola D., 2001 Inferring the trajectory of genetic variance in the
Stam P., 1980 The distribution of the fraction of the genome identical by descent in finite
Strandén I., Christensen O. F., 2011 Allele coding in genomic evaluation. Genet Sel Evol 43:
25.
Sved J., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite
Tsuruta S., Misztal I., Lawlor T. J., 2004 Genetic Correlations Among Production, Body Size,
36.
VanRaden P. M., 2008 Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91:
4414–4423.
Wang H., Misztal I., Legarra A., 2014 Differences between genomic-based and
pedigree-based relationships in a chicken population, as a function of quality control