Long-distance associations generate erosion of genomic breeding values of candidates for selection

D Boichard, S Fritz, P Croiseau, V Ducrocq, B Cuyabano, T Tribout

Long-distance associations generate erosion of genomic breeding values of candidates for selection

D. Boichard[1], S. Fritz[1,2*], P. Croiseau[1], V. Ducrocq[1], B. Cuyabano[1], and T. Tribout[1]

[1]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France;
[2]Allice, 75012 Paris, France. sebastien.fritz@inrae.fr

Abstract.
Most validation studies of genomic evaluation observe inflation, *i.e.* regression coefficients of the later phenotypes on early predictions smaller than one. In this paper, we show that this is due to non-zero contributions of distant QTLs, especially those located in other chromosomes than the SNPs. These effects result from linkage disequilibrium (here measured with r^2) present in the reference data but eroded in the population of candidates. In six French dairy cattle breeds average r^2 is low across chromosomes but a substantial proportion of SNP pairs show r values higher than 0.05. A simulation study based on real genotypes from the Normande breed shows that 5% of the SNP effect's variance is explained by distant QTL with associations not maintained across generations. A real Holstein example illustrates extensive across chromosomes covariances between effects. It is thus recommended to erode SNP effects to compute unbiased genomic values of candidates to selection.

Introduction

In genomic evaluation, SNP effects are estimated in a reference sample and applied to selection candidates. This method is extensively used to select candidates at an early stage of their life or yet without phenotypic information. The standard interpretation is that a trait depends on QTLs, and those genetic markers in close LD to these QTL are good proxies for them. Implicitly, this assumes that estimated marker effects reflect those of the neighbouring QTL. Under this assumption, associations observed in the reference sample should be very similar in the next generation as short-distance LD erodes slowly due to recombination.

It is, however, well known that genomic evaluation efficiency is highly dependent on the relatively close relationship of the candidates to the reference sample (Habier et al, 2007, 2013; Legarra et al, 2008; Pszczola et al, 2012). Many studies have shown the limited gain in accuracy in multi-breed evaluation (Erbe et al, 2012; Hozé et al, 2014), illustrating that distant reference data are not informative. Other studies have shown decrease in accuracy over generations when the reference sample is not updated (Sonneson et al, 2009; Solberg et al, 2009). Moreover, it has been observed that the absence of the parents in the reference sample directly influences the prediction accuracy of the selection candidates. All these results suggest that SNP effects erode as the distance between candidates and reference sample increases.

Despite these observations, validation studies of genomic evaluations are generally based on the regression of later performances on the early predictions. These studies frequently observe an inflation pattern, *i.e.* the regression coefficient is systematically below 1, meaning that later performances of the best candidates were below those initially predicted (and later performances, if any, of the worst candidates were above those initially predicted).

An interpretation is that some long-range LD exists, even between chromosomes, and, accordingly, many markers may capture some partial effects of supposedly unlinked QTL. Long-distance LD is much lower than short-distance LD, but the number of variants involved is much higher and their joint effects can explain a substantial proportion of the predicted genomic value. In this paper, we show that markers capture some effects of distant QTL due to long-distance LD. Because this long-range LD dissipates over generations, the erosion of marker effects must be taken into account to predict the genomic value of the candidates.

Material and Methods

LD across chromosomes was measured in the female reference data of six French dairy cattle breeds (Holstein, Montbeliarde, Normande, Abondance, Tarentaise, Vosgienne). The size of these reference samples varied from 2617 to 362,363. Vosgienne, Tarentaise, and Abondance are regional mountain breeds, Montbeliarde and Normande are large national populations (18% and 7% of the French dairy herd, respectively) and Holstein (70%) is international. One every 20 SNP was considered, resulting in a sample of ~3 million r^2 values for each breed (vs ~1.2 billion in total).

To study the impact of long-distance LD on SNP effects, we used the Normande population (N=69,220 genotyped cows with records). Only the first five chromosomes (ns=13,608 SNP) were considered. Two hundred (nq=200) additive QTLs were sampled at random among those SNPs with MAF>0.02 over the first four chromosomes whereas chromosome 5 remained empty of QTLs. Additive QTL effects were independently drawn from a normal distribution assuming a heritability of 0.3. Two scenarios were tested: (1) the SNP-BLUP model accounted for the 13,608 SNPs including the QTL, a situation believed to minimize the impact of long-distance LD; (2) the SNP-BLUP model did not include the QTL, a situation in which QTL effects are to be distributed on more markers. The contribution of each QTL to each SNP was estimated as follows. The SNP-BLUP equations can be written as $(\mathbf{M'M} + \lambda\ \mathbf{I})\ \hat{\boldsymbol{s}} = \mathbf{M'}\ \mathbf{y}$, with $\mathbf{M}$ the (N x ns) matrix of centered and scaled genotypes, $\mathbf{s}$ the vector of SNP effects, $\mathbf{y}$ the vector of phenotypes adjusted for fixed effects, and $\lambda = \sigma_e^2 / \sigma_s^2$. According to the simulation, the phenotype can be written as $\mathbf{y} = \mathbf{Pq} + \mathbf{e}$, *i.e.* the sum of nq QTL effects and an error term, with $\mathbf{P}$ the (N x nq) matrix of genotypes at the QTL level and $\mathbf{q}$ the vector of true QTL effects. Therefore, the equations can be rewritten as $\hat{\boldsymbol{s}} = (\mathbf{M'M} + \lambda\ \mathbf{I})^{-1}\ \mathbf{M'(Pq + e)}$. If $\mathbf{c_i}$ is line i (corresponding to SNP i) of $\mathbf{C} = (\mathbf{M'M} + \lambda\ \mathbf{I})^{-1}$, the contribution of QTL j to SNP effect i is $\mathbf{c_i}\ \mathbf{M'}\ \mathbf{P_j}\ q_j$. These contributions were squared and summed within each of the 6 following categories of distance between QTL and SNP: (1) the QTL is the SNP; (2) the distance between the QTL and the SNP is less than 5 Mb; (3) this distance is between 5 and 20 Mb; (4) the QTL and the SNP are on the same chromosome but more than 20 Mb apart; (5) the QTL and the SNP are on different chromosomes; category (6) corresponded to the special case of SNPs on chromosome 5 with no QTL. Thirty replicates were simulated.

This way to estimate contributions by accumulating squared effects by individual SNP does not account for the covariances between SNP effects. This method is therefore only an approximation (de los Campos et al, 2015). However, the long-distance covariances have a major impact and reinforce the contribution of distant markers. They are illustrated by the part of DGV variance explained by haplotypes of increasing size. These results were from the French Holstein evaluation for resistance to paratuberculosis (Sanchez et al, 2022, WCGALP).

Results

Table 1 presents different LD statistics across chromosomes in the reference samples of six French dairy cattle breeds. Average r^2 values are small, which at first glance may suggest that across chromosomes effects are negligible. These values decreased when the size of the breed (reference sample, number of females in the breed, or effective size) increased. Note, however, the parameter of interest here is r instead of r^2, as the impact of a QTL on a SNP effect is proportional to r. The proportion of SNP pairs with |r| greater than 5% varied from 1.5% to 33%, depending on the breed. The larger the reference sample, the lower this proportion. Nevertheless, even with a few percentuals, one to several thousand SNP have some correlations to QTLs (assuming that these r distributions are the same between SNP and QTL).

Table 2 presents the relative contribution of the 6 classes of QTL-SNP distance, expressed in proportion of the total sum of squared SNP effects. In the model including the causal variants, these variants succeeded to capture the largest part of the genetic variability, leaving little

contributions to the nearby markers. Contributions of distant QTLs were nearly zero. In the model without causal variants, more similar to a real situation, the distribution of contributions was spread on more distant markers, even across chromosomes, although the largest part of the variability was still captured by the closest markers. This result was obtained with only five chromosomes and the contribution to distant markers is likely to be even larger with 29 chromosomes, due to the multiplicity of the across chromosomes contributions.

Table 1. Statistics of |r| and r^2 values across the 29 chromosomes in reference samples of six French dairy cattle breeds (selection of 1 every 20 SNP within chromosome).

Breeds	# cows in reference sample	Mean(|r|) *	% |r| > 0.03	% |r| > 0.05	Mean(r^2)
Vosgienne	2617	0.0420	54	33	0.0029
Tarentaise	3788	0.0225	41	18	0.0015
Abondance	7115	0.0268	35	15	0.0012
Normande	69,220	0.0206	25	7	0.00073
Montbeliarde	185,053	0.0173	18	4	0.00053
Holstein	362,363	0.0148	12	1.5	0.00038

* statistics based on 2,812,741 to 3,231,80 r-values per breed

Table 2. Relative contribution of each of the 6 classes of QTL-SNP pairs defined according to their distance. The contribution of a class is the sum of squared effects. Results on 30 replicates.

Classes of QTL-SNP relationship	Relative contribution (%)		# of QTL contributions to SNP
	Model including causal variants	Model without causal variants	
1: QTL=SNP	95.2	0.0	200
2: distance(QTL,SNP) < 5 Mb	4.6	73.8	42,660 +/- 530
3: 5 Mb < distance < 20 Mb	0.1	19.6	112,500 +/- 1500
4: distance(QTL,SNP) > 20 Mb	0.02	0.6	410,000 +/- 3700
5: QTL and SNP on different chromosomes	0.02	2.4	1,671,000 +/- 4500
6: SNP on chromosome 5 (without QTL)	0.02	3.5	485,200

These contributions at long-distance are the consequences of long-distance LD (**M'P** is a measure of the covariance between SNP and QTL). This basal LD is a function of the effective size Ne and remains more or less constant across generations on average at the genome level. Two opposite forces explain this steady-state LD: (1) The existing long-range LD is reduced by a 0.5 factor at each generation due to segregations; and (2) it is regenerated by drift (and possibly selection) at random. However, this is not true for a given pair of loci and the LD between 2 unlinked loci existing in the reference sample is reduced by a 0.5 factor.

Table 3. Proportion of DGV explained by haplotypes of various sizes in French Holstein

Interval length (# markers)	1 (single markers)	10	100	500	Entire Chromosomes	Whole-genome
% variance explained	1	8.5	26	47	58	100
# intervals	53,469	5,360	548	121	29	1

Variances were computed from partial DGV per interval in the evaluated population

Finally, Table 3 illustrates that the part of the DGV variability explained was very small with individual variants and gradually increased with haplotype size, showing the huge amount of positive covariances between marker effects. Even when entire chromosomes were considered, the sum of the 29 variances represented only 58% of the total, showing that 42% of the total variance originated from inter-chromosomal covariances.

Conclusions

The practical consequences are important. In the overall prediction of candidates, the short distance contributions are rather stable, slightly eroded by recombinations (Dekkers et al, 2021), but the long-distance contributions are divided by 2 at each generation. According to the relative weight of short and long-distance LD, the overall erosion may strongly fluctuate, but it is always important. Assuming this erosion factor per generation (ρ) is known, eroded direct genomic values (DGV) of candidate i can be estimated from the raw DGV of i and of its parents s and d: $DGV_i = 0.5\ (DGV_s + DGV_d) + (1- \rho)^k\ MS_i$, with MS being the Mendelian sampling component and k the number of generations between i and the reference sample on both paternal and maternal pathways. This formula is recursive, as sire's and/or dam's DGV should also be eroded if they do not belong to the reference.

Further investigation is needed to estimate this ρ erosion factor. It of course depends on basic LD in the population (*i.e.* on Ne and the genome length L). But it also probably depends on the genetic architecture of the traits (number and magnitude of QTL effects) and the model used (SNP-BLUP/GBLUP vs Bayesian models). However, one can anticipate that: (1) Inflation factors ($1 - \hat{b}$) frequently observed are between 0.1 and 0.2 and give an idea of the importance of the phenomenon; (2) Models including causal variants are more persistent and less subject to erosion; (3) Breeding schemes with accelerated generations without updating reference data accumulate more erosion; (4) Models including a residual polygenic component present less inflated predictions because they combine two estimates of the MS term, the genomic one which is inflated and a polygenic one which is equal to zero. We believe that accounting for erosion is more rigorous and accurate than adding an arbitrary polygenic effect.

References

Dekkers J.C.M., Su H.L., and Cheng J. (2021) Genet Sel Evol 53, 81. https://doi.org/10.1186/s12711-021-00675-6

de los Campos G., Sorensen D., and Gianola D. (2015) PLoS Genet 11, e1005048. https://doi.org/10.1371/journal.pgen.1005048

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., et al (2012) J Dairy Sci 95(7), 4114-4129. https://doi.org/10.3168/jds.2011-5019

Habier D., Fernando R.L., and Dekkers J.C.M. (2007). Genetics 177, 2389-2397. https://doi.org/10.1534/genetics.107.081190

Habier D., Fernando R.L., and Garrick D.J. (2013) Genetics, 194, 597–607, https://doi.org/10.1534/genetics.113.152207.

Hoze C., Fritz S., Phocas F., Boichard D., Ducrocq V., et al (2014) J Dairy Sci 97(6), 3918-3929. https://doi.org/10.3168/jds.2013-7761

Legarra A., Robert-Granie C., Manfredi E., and Elsen J.M. (2008) Genetics 180(1), 611-618. https://doi.org/10.1534/genetics.108.088575

Pszczola M., Strabel T., Mulder H.A., and Calus M.P.L. (2012) J Dairy Sci 95(1), 389-400. https://doi.org/10.3168/jds.2011-4338

Solberg T R., Sonesson A.K., Woolliams J.A., Odegard J., and Meuwissen T.H.E. (2009) Genet Sel Evol 41, 53. https://doi.org/10.1186/1297-9686-41-53

Sonesson A.K., Meuwissen T.H.E. (2009) Genet Sel Evol 41, 37. https://doi.org/10.1186/1297-9686-41-37