Use of whole sequence GWAS to improve genomic evaluation in dairy cattle

Thierry Tribout, D Boichard, M P Sanchez, Sebastien Fritz, Pascal Croiseau

Use of whole sequence GWAS to improve genomic evaluation in dairy cattle

P. Croiseau[1], T. Tribout[1], D. Boichard[1], M.P. Sanchez[1] & S. Fritz[2]

[1] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France
pascal.croiseau@inra.fr (Corresponding Author)
[2] Allice, F-75012 Paris, France

Introduction

Most routine genomic evaluations in dairy cattle worldwide use a 50K SNP chip. This chip has several advantages such as good coverage of the genome or high minor allele frequencies (MAF) that makes it suitable for most dairy cattle breeds. Nevertheless, including causal variants in genomic prediction, rather than 50K SNP in linkage disequilibrium (LD) with causal variants, may improve evaluation accuracy. Recently, the "1000 bull genomes" population has made it possible to impute whole genome sequences for a large number of animals. Even if imputation accuracy is still limited for rare variants (Brondum et al., 2014), it becomes possible to directly pinpoint the main causal variants, or at least variants in very high LD with them. For computational reasons, whole genome sequence data cannot be used in routine genomic evaluation and only the most predictive part of them should be identified before any use.
In this study, we propose to construct virtual chips containing 50k variants (V50K) selected from whole genome sequence, therefore compatible with a routine use. In this aim, GWAS were performed at the sequence level. However, valorizing GWAS results is not straightforward. Indeed the selection of putative variants depends on MAF constraints, on p-value thresholds, on the weight put on functional annotation, or on constraints on the minimum distance between candidate variants. Therefore, several scenarios were tested to select these 50k variants. These strategies, based on the above parameters, fall into two categories: either chips for which all SNPs are selected from sequence data; or chips for which only a fraction of the SNPs of the original Illumina Bovine chip SNP50 Beadchip® (IB50K) are replaced. This study presents the results of these different strategies applied to the medium size Montbeliarde population.

Material and methods

Data

A total of 3,130 bulls from the French Montbéliarde breeds were studied for a panel of ten traits related to milk production (milk, fat and protein yields, fat and protein contents), cow fertility (conception rate at each service), conformation (height at sacrum and rear udder width) and functional traits (milking speed and somatic cell score). The phenotypes of bulls were Daughter Yield Deviations (DYD, VanRaden and Wiggans, 1991). All bulls were genotyped on the IB50K and imputed at the sequence level using FIMPUTE software (Sargolzaei et al., 2014) with a two-step procedure described by Van Binsbergen et al. (2014). IB50K genotypes were first imputed to high-density with a Montbéliarde reference population made of 521 key ancestors genotyped on the BovineHD BeadChip® (IBHD). IBHD genotypes were then imputed to sequence with the multi-breed 1,000 bull genomes reference

population (1,124 sequenced bulls of the Run 4). The population was then divided in training and validation sets including 2,608 and 522 bulls, respectively.

GWAS analyses

GWAS analyses were carried out at the sequence level on the training population using a mixed model as implemented in Wombat (Meyer et al., 2007; 2012). DYDs were the dependent variables and individual breeding value was included as a random effect with relationships among animals accounted for via the pedigree. To account for their heterogeneous variances, DYDs were weighted by their number of effective records (Fikse and Banos, 2001). Allele dosage for each SNP was included in the model as a fixed effect. Significance level for each SNP was calculated from the resulting t-statistic assuming a two-tailed t-distribution. Table 1 presents the number of significant SNP for protein yield as an example, at different thresholds and according to MAF of the SNP. The GWAS results were used to select markers and design V50K chips.

Table 1. Number of significant SNP (at different –log(p-value) thresholds) for protein yield according to the Minor Allele Frequency.

		-log(p_value) threshold					
		[1.3 - 3]	[3 - 4]	[4 - 5]	[5 - 7]	[7 - 10]	>10
MAF	> 0.3	702 378	89 680	28 278	15 613	1 852	1
	[0.1 , 0.3]	883 871	122 681	45 358	24 599	4 385	228
	[0.05 , 0.3]	315 185	52 523	21 184	11 506	2 806	166
	[0.01 , 0.05]	371 893	85 186	39 114	32 791	8 641	1 658
	[0 , 0.01]	119 683	28 738	25 011	29 864	18 234	6 579

Strategies to build V50K

Two general strategies were explored. In the first strategy, all the SNPs of the IB50K chip were replaced by SNPs from the sequence. In the second one, only a part of the SNPs of the IB50K chip was replaced whereas a number of the original IB50K markers were kept. The total number of markers of all V50K was 43,800, corresponding to the actual number of markers of the IB50K chip used after quality controls.

Although this is not realistic in practice, theV50K chips were designed within trait to obtain the highest efficienty. Five scenarios were tested for each trait, for a total of 50 virtual chips.

Strategy 1, three scenarios:

- PEAK: a selected SNP has the lowest p-value in a window of ±150 SNP. Only the 43,800 most significant ones were retained.
- COVER: The genome was divided in 43,800 segments. In each segment, the SNP with a MAF higher than 1% and the lowest p-value was retained.
- COVER2: Same as COVER but in addition to p-value and MAF conditions, the SNP was required to be located in a gene (if any).

Strategy 2, two scenarios:

- OPT_QTL: The genome was split in 1Mb-intervals. Candidate SNPs were required to fulfil conditions on MAF (3 values tested: 1, 3 and 5%) and significance level (3 p-value thresholds tested : 10^{-3}, 10^{-4} and 10^{-5}). All intervals with at least one candidate SNP were retained. In each of these regions, the most significant SNP replaced the IB50K SNP with the lowest MAF. Using these scenarios, between 169 and 747 SNP were replaced.
- Bottom-up: GBLUP and BayesCPi analyses were performed on the IB50K data of the training set; the SNP of the IB50K chip with estimated effects equal or very close to 0

were removed and replaced by the sequence SNP with the lowest p-value in a region of ±0.5Mb. Using this scenario, around 3000 SNP were replaced.

Genomic evaluation procedure

IB50K and all V50K chips were used in a genomic evaluation procedure for the 10 selected traits. The method used was a BayesCπ, as implemented in GS3 software (Legarra et al., 2011) with the following model:

$$\mathbf{y_i} = \boldsymbol{\mu} + \sum_{\mathbf{j=1}}^{\mathbf{43800}} \boldsymbol{\beta_j x_{i,j} \delta_j} + \mathbf{u_i} + \mathbf{e_i} \quad (1)$$

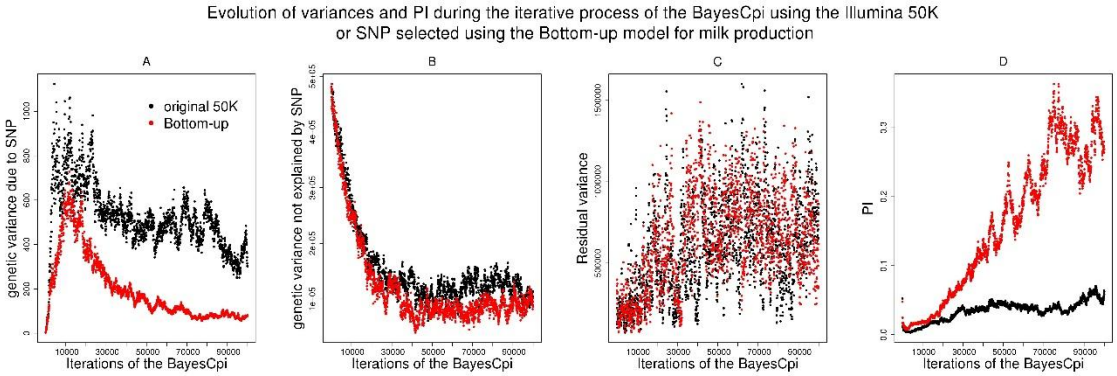
where y_i is the performance of animal i, μ is the mean, x_i is the copy number of the second allele of marker j in animal i, β_j is the allelic substitution effect for marker j, δ_j is a 0/1 variable indicating whether marker j has a non-zero effect, u is a random vector polygenic effects for animal i and e_i is the random error term for animal i.

GEBV accuracy was estimated through the weighted correlations between DYD and GEBV in the validation set, where the weight corresponds to the sire's number of effective records.

Results and discussion

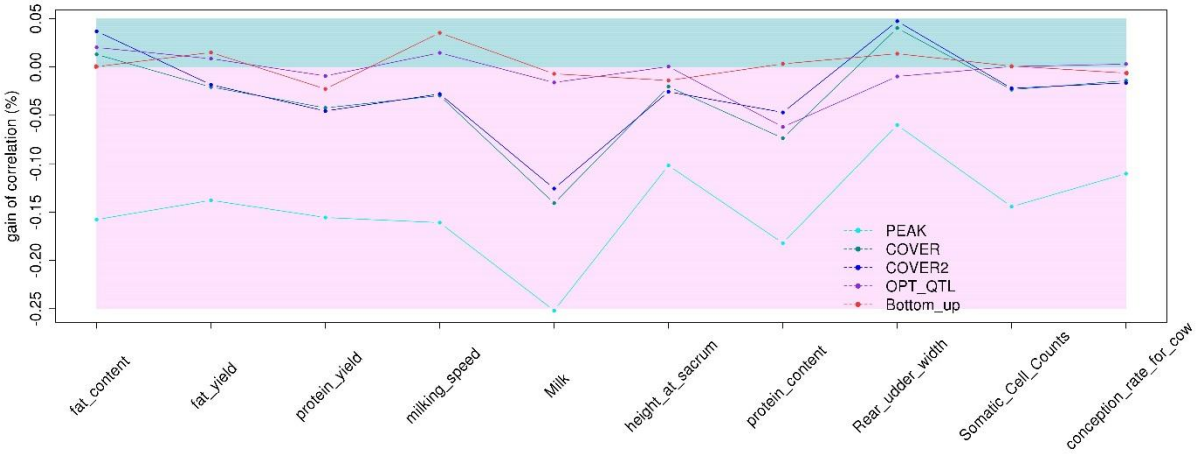
The BayesCπ method estimates the polygenic variance, the SNP variance, the residual variance, and the probability of inclusion of each SNP in the model. Figure 1 presents the evolution of these variances and probability of inclusion during the BayesCπ iterative process with the IB50K chip or with the Bottom-up scenario for milk yield. A positive impact of the use of virtual chips on genetic parameters estimation can be noted through a better convergence of the SNP variance (A), a lower genetic variance not explained by SNP (B) which means that a higher proportion of the genetic variance is explained by the SNP, and a higher proportion of retained SNP at each iteration (D). The latter observation means a higher number of SNP influencing the trait, which is expected from SNP selection. However, because the part of the genetic variance explained by the SNP is not much increased, the SNP variance tended to decrease (A). A similar behaviour was observed with COVER and OPT_QTL model and for other traits.

Figure 1. Evolution of genetic variance explained by the markers (average per retained marker) and not explained by the markers, of residual variance, and of probability of inclusion during the BayesCπ iterations using the IB50K chip or with the Bottom-up scenario for milk yield.



Compared to IB50K, a loss or a low gain of accuracies of genomic predictions was observed with V50K (Figure 2). The PEAK strategy provided the worse results with an average loss of relative accuracy of 15%. With COVER and COVER2 strategies, a loss of correlation is observed for 8 traits over the 10, but a gain is observed for fat content (+4% with COVER2) and rear udder width (+4 and +5% with COVER and COVER2, respectively). With OPT_QTL and Bottom-up strategies, a gain of correlation is observed for 6 (OPT_QTL) and 7 (Bottom-up) traits but this gain is limited to 1 to 4%, respectively. Finally, for 3 traits (protein yield, milk yield, height at sacrum), no improvement was observed whatever the V50K used.

Figure 2. Variation of correlation between DYD and GEBV with V50K chip vs IB50K.



Conclusion

Some of the tested strategies resulted in an increase in prediction accuracy compared to IB50K. However, this increase appears limited, and no improvement was obtained for 3 traits over of the 10 considered, despite the V50K were designed specifically for each trait, which should have been a favourable and optimistic situation.

One explanation of this limited improvement could be the low number of sequenced Montbéliard bulls in the 1,000 bull genome Run 4 reference population, which possibly results in a limited imputation accuracy for rare variants. This analysis will be repeated in Holstein to test this hypothesis. The persistence of LD over long distances in purebred

populations, allowing the markers of the IB50K chip to capture most of the effects of causal variants, could be another explanation.

List of References

Brøndum, R. F., Guldbrandtsen, B., Sahana, G., Lund, M. S., & Su, G, 2002. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics, 15, 728.

VanRaden, P. M., & Wiggans, G. R, 1991. Derivation, calculation, and use of national animal model information. Journal of Dairy Science, 74, 2737-2746.

Fikse, W. F., & Banos, G, 2001. Weighting factors of sire daughter information in international genetic evaluations. Journal of Dairy Science, 84, 1759-1767.

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S, 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics, 15, 478.

Van Binsbergen, R., Bink, M. C., Calus, M. P., Van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., & Veerkamp, R. F, 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution, 46, 41.

Meyer, K, 2007. WOMBAT- A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University-Science B, 8, 815-821.

Meyer, K., & Tier, B, 2012. "SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model. Genetics, 190, 275-277.

Legarra, A., A. Ricard & O. Filangi, 2011. GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCπ), http://snp.toulouse.inra.fr/~alegarra/