Technical options for all-breed Single-step GBLUP for US dairy cattle

A. Legarra[1,2], M. Bermann[2] P.M. VanRaden[3], E.L. Nicolazzi[1], R.R. Mota[1], J.M. Tabet[2], D.L Lourenco[2] and I. Misztal[2]

[1] Council on Dairy Cattle Breeding, 4201 Northview Drive, 20716 Bowie MD, USA
[2] Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
[3] U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705-2350, USA
Corresponding author: andres.legarra@uscdcb.com

Abstract

The multi-step method for genomic prediction has worked remarkably well for US dairy cattle, but intense genomic selection makes recent genetic trends difficult to estimate in pedigree-only based BLUP evaluations. Thus, the introduction of routine single-step GBLUP (ssGBLUP) is under study. The large size of US dairy cattle data precludes naïve approaches for genomic prediction. Here we present the technical choices and needs of an all-breed (6 breeds and all existing crosses), ssGBLUP applied to different sets of traits within trait groups such as fertility, livability and health data. For each trait group, first, we prune pedigree to animals with records and their ancestors, reducing the size of pedigree and improving memory use and convergence. The model includes only genotypes of animals in this pruned pedigree, and we predict the other animals later either using Parent Average (if not genotyped) or sum of SNP effects (if genotyped). The set of markers is the usual CDCB set with 78,964 markers and included autosomes and sex chromosomes. The method for ssGBLUP was G-matrix with Algorithm for Proven and Young (APY) with metafounders (MF). APY largely reduces computational needs whereas MF provides smooth solutions for unknown origins and automatic compatibility of pedigree and genomic relationships within and across breeds. The gamma matrix was constructed based on base allele frequencies across breeds and increases of inbreeding within breeds. Core animals were chosen within breed, in a heuristic but complete and repeatable manner: genotyped sires with more than a certain number of daughters in records, and a deterministic subset of genotyped cows with records. This resulted in ~45K animals in the core and ~2M non-core animals for fertility evaluations. Still memory needs are large as G_APY inverse, stored in double precision, takes ~720 Gb. Thus, we used memory mapping (mmap) to assign memory to disk space. For the case of fertility (4 traits), computation of G-1_APY took 28h and 100 Gb of RAM using mmap. Solving MME took 22h, 120 Gb of RAM and 476 rounds of PCG. Genomic reliabilities took 120 Gb of RAM and 8h per trait. Backsolving for SNP solutions took negligible time and memory. Owing to the developments reported here, computations for ssGBLUP in this very large database can be done with reasonable time and memory.

Key words: metafounders, memory mapping, pedigree, genomic

Introduction

Genomic predictions in dairy cattle started with quite simple multi-step methods consisting in traditional pedigree-based evaluations followed by genomic predictions based on de-regressed proofs of the reference population – those animals with genotypes and some sort of information from traditional BLUP. However, multi-step methods do not use all available information and, probably more important, traditional evaluations produce biased genetic trends. Single-step methods (either in SNP-BLUP or GBLUP flavors) can instead use all information to estimate unbiased trends and improve reliability.

Therefore, national dairy cattle evaluations are gradually shifting to single-step methods. Single-step methods are complex for two reasons. First, the elementary values handled are orders of magnitude larger than pedigree-based evaluations. For instance, a genetic evaluation with 1 million animals in pedigree uses a pedigree list of 3 million points. The same animals in a pure genomic evaluation would use 50 billion points: 50K (SNPs) times 1 million (cows). The second reason for the complexity is the easy algebra but complex operations in the single-step methods.

The US genetic evaluation system at Council on Dairy Cattle Breeding (CDCB) is very large, including roughly 60 million animals with records, 100 million animals in pedigree, more than 8 million animals genotyped and 50 traits grouped in different models. CDCB, AGIL (USDA) and University of Georgia are testing single-step methods using the blupf90 suite of programs. This led us to define technical options to avoid the use of very large resources (time, memory, disk space) or extensive reprogramming. We present these technical options here as they might be of interest for other practitioners.

Materials and Methods

Pruning pedigree and markers

The CDCB evaluates several trait groups (yield, somatic cell score, livability, productive life, fertility, gestation length, health, residual feed intake (RFI), heifer livability, calving ease and type traits) including a total of 50 traits – see https://uscdcb.com/individual-traits/ . Residual feed intake is a Holstein-only evaluation; type traits are separate purebred evaluations; the rest are all-breed evaluations. The number of animals with phenotypes varies enormously from ~8K for residual feed intake to 40M for yield traits. There are at this moment (June 2024) 9 million genotyped animals, all imputed to 79K SNPs. However not all this information is needed for the genomic evaluation itself. The CDCB receives pedigrees and genotypes for animals that are not directly related to the evaluations – because they are foreign animals or because they belong to herds that do not contribute information. They are related to records through pedigree, genotypes, or both.

Consider pedigree first. The set of animals in records for yield (the trait with largest database) and its ancestors constitute 60M animals. The set of animals in records for residual feed intake (the trait with smallest database) and its ancestors constitute roughly 60K animals. Although in theory one could include all 100M animals in the Mixed Model Equations (MME) for all traits, this is clearly an overkill. Preliminary analyses using the blupf90 family showed that solving of the Mixed Model Equations with all 100M animals in pedigree needed stricter convergence criteria (as some animals are very distantly related to phenotypes) than the trimmed 60M pedigree. Therefore we trim the pedigree, solve the MME, and then predict the trimmed animals by pedigree relationships (Henderson, 1977). This is done via Parent Average from oldest to youngest in the trimmed animals.

Then consider genotypes. One way of understanding single-step is that it improves pedigree relationships of non-genotyped animals via related genotyped animals. Thus, and contrary to pedigree BLUP, a young animal with no phenotype and no progeny with phenotype may contribute to improve the elements in **H** for its non-genotyped parent(s). However, it is commonly accepted that this improvement is very small. Thus, we decided to retain genotypes of animals directly related to phenotypes: animals with records and ancestors, reducing the number of genotypes from 9M to 2M for traits such as yield.

In other words, first we built a pedigree consisting of animals in records and all ancestors; then, we extracted the genotypes of animals included in this subset pedigree. The GEBVs of the remaining animals can be predicted based on SNP effects and pedigree predictions (e.g., Vandenplas et al., 2023).

Metafounders

To model missing parentship and different breeds levels we fit metafounders defined by breed, year of birth, and selection path. Metafounders give smoother estimates (Legarra and VanRaden, 2023) and compatibility with genomic relationships (Legarra et al., 2015). Within trait group, we defined a joining strategy that first compacts the definitions "forward" until first phenotypes appear (e.g. for health traits) and then "backwards" to achieve a minimum "pseudo-count of records in progeny" per level of metafounders. These results in varying numbers of metafounders levels per trait group, up to approximately 300 at most. The relationship matrix across metafounders was obtained using base allele frequencies estimated from old genotypes in the database, plus a strategy using increase of inbreeding for more recent ones (Legarra et al., 2024). A "heatmap" of metafounders relationships is in Figure 1.

Using the algorithm for Proven and Young

For these tests we decided to use the Algorithm for Proven and Young, so called APY, which uses a sparse representation, $\mathbf{G}_{APY}^{-1}$, of the conditional covariances across individuals (Misztal et al., 2014; Misztal, 2016). It can also be seen as an approximate sparse inverse of the genomic relationship matrix. The APY algorithm has several advantages: it is fast and memory wise, easy to program. However, it requires the definition of a set of "core" animals representing the whole population. This was done using some ideas from Cesarani et al. (2022) and some new ones. We also wanted (1) to have both bulls and cows and (2) to avoid randomness, because it makes troubleshooting genetic evaluations more complex. The choice of core was done by breed as follows.

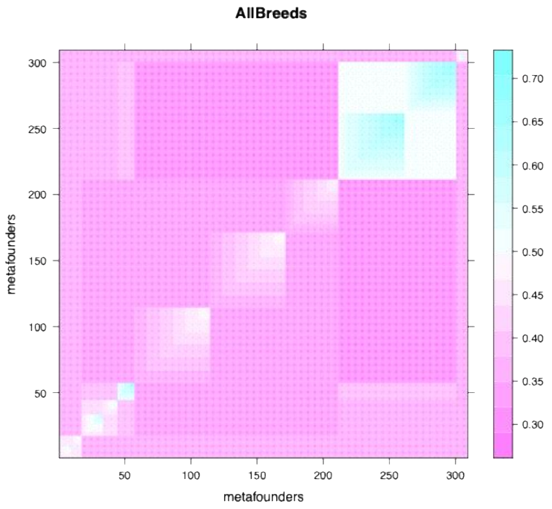

Figure 1. Gamma relationship for metafounders, sorted by breed and pedigree path

The genotypes that we used for large trait groups as fertility or yield consist in a very large number of Holsteins (almost 2M), a medium number of Jerseys (300K), a small number of crossbred animals (called "XX") (50K) and smaller number (<10K) for each of Ayrshire, Brown Swiss and Guernsey. At this stage it is unclear if XX genotypes will be included in the possible "routine" single-step, but we did that to test the most complex case. Very old genotyped animals (<1990) were *not* chosen as core as they are not truly representative of their respective periods. For AY and GU, all animals were included as core. For BS and XX 5K animals were needed as core, and 15K for JE and HO. These numbers have been found in previous studies – see Cesarani et al. (2022).

Then, within population: first, we first chose all genotyped sires with >100 daughters (Jersey, XX) or >500 (Holstein) with records. This left some spots to fill, that were filled with cows with records using a deterministic function module(anim_key,n) where anim_key is the unique integer used at CDCB for identification and n is a number to fill in the empty spots in the core. Table 1 gives an overview of the final numbers.

Table 1: number of animals and core genotypes for tests on livability

Breed	Genotypes	Core needed	Sires flagged	Cows flagged
Ayrshire	1,608	(all)	311	1175
Brown Swiss	9,560	5K	611	4313
Guernsey	3,561	(all)	219	3258
Holstein	1,669,795	15K	6890	8113
Jersey	300,976	15K	3186	11883
Crosses	56,528	5K	141	4616

Memory mapping

Even with APY, $\mathbf{G}_{APY}^{-1}$ stored in double precision requires ~700 GB for 45K animals in core and 2M animals in non-core. This is still less than the matrix of genotypes in double precision for the ~2M animals and 79K genotypes. This matrix is first formed by cross-products of blocks with program preGSf90, blended (5% or 10%) with a residual polygenic relationship matrix $\mathbf{A}_{(\Gamma)22}$ based in pedigree, then inverted. Allele frequencies are fixed to 0.5 as assumed by theory of metafounders. Then $\mathbf{G}_{APY}^{-1}$ is used by program blup90iod3 (solving the MME) and accGS2f90 (accuracies as in Bermann et al., 2022a).

The iterative method by Preconditioned Conjugate Gradients in blup90iod3 essentially consist in multiplications of the MME times a vector of solutions. This has a low cost for the pedigree + pedigree relationships part, which in addition can be easily solved by iteration on data algorithms. However, the contributions of $\mathbf{G}_{APY}^{-1}$ to the MME is more expensive if handled in memory. To alleviate memory needs, we used the programming technique called memory mapping (mmap) (https://en.wikipedia.org/wiki/Mmap) which allows mapping memory to disk space. Using this technique, RAM is reduced to 120Gb instead of 700Gb.

Backsolving for SNP solutions

The SNP effects estimates are needed for Indirect Predictions of animals not included explicitly in the MME, and also for new animals arriving to the database in between full runs. The SNP effect estimates can be obtained backsolving from GEBVs of core animals obtained in the full run (Bermann, 2022b). This has low computation cost as the core animals are a reduced number and $\mathbf{G}_{core,core}^{-1}$ is already available as part of $\mathbf{G}_{APY}^{-1}$.

Rough timings and memory

Here we give some crude numbers. We have made different tests across different servers, trait groups and options. The examples are for fertility traits: four traits, low heritability, with records dating back to 1960 – see Legarra and VanRaden (2023) for a more complete description.

There are roughly 100M lactation records belonging to 40M animals, with different patterns of missing values and different models across traits. The pedigree including animals in records and ancestors contains 60M animals. Of these, 2M animals are genotyped and their genotypes included in the prediction, and of these, 45K constitute the core, in a manner similar to Table 1.

We used 16 threads. Preparing $\mathbf{G}_{APY}^{-1}$ with preGSf90 took 16h and 720Gb of RAM or 28h and 120 Gb of RAM using mmap. For all the next operations we used mmap. Solving MME by blup90iod3 took 22h, 120 Gb of RAM and 476 rounds of PCG. Genomic reliabilities using an approximation to the inverse of the MME (Bermann et al., 2002a) took 120 Gb of RAM and 8h per trait. Backsolving for SNP solutions took negligible time and memory. These numbers are very similar to Cesarani et al., 2022.

Conclusions

Testing single-step forces to make explicit the choices and steps of the genetic evaluation systems and pipelines. The choices that we present here adapt easily to the diverse variety of information, traits and population at CDCB, while they should guarantee a fair, unbiased evaluation on time without using extensive computing resources.

Correct handling of missing pedigree and different breeds, e.g., using metafounder, is important for unbiased results. Choices of core and trimming pedigree are essential to save memory and computing time.

Acknowledgments

AGIL staff and USDA funding of project 8042-31000-113-000-D, “Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals”, and UGA funding Agriculture and Food Research Initiative Competitive Grant no. 2020-67015-31030 from USDA

References

Bermann, M., D. Lourenco, and I. Misztal. 2022a. Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young. *Journal of Animal Science* 100:skab353. doi:10.1093/jas/skab353.

Bermann, M., D. Lourenco, N.S. Forneris, A. Legarra, and I. Misztal. 2022b. On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young. *Genet. Sel. Evol.* 54:52. doi:10.1186/s12711-022-00741-7.

Cesarani, A., D. Lourenco, S. Tsuruta, A. Legarra, E. L. Nicolazzi, P. M. VanRaden,, and I. Misztal. 2022. Multibreed genomic evaluation for production traits of dairy cattle in the US using single-step GBLUP. *J. Dairy Sci.* 105:5141-5152. doi.org/10.3168/jds.2021-21505.

Henderson, C. 1977. Best linear unbiased prediction of breeding values not in the model for records. *J. Dairy Sci.* 60:783–787.

Legarra, A., O.F. Christensen, Z.G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* 200:455–468. doi:10.1534/genetics.115.177014.

Legarra, A., and P.M. VanRaden. 2023. Effect of modelling unknown parent groups and metafounders on the historical genetic trend of fertility traits. *Interbull Bulletin* 59:11–14.

Legarra, A., M. Bermann, Q. Mei, and O.F. Christensen. 2024. Estimating genomic relationships of metafounders across and within breeds using maximum likelihood, pseudo-expectation–maximization maximum likelihood and increase of relationships. *Genet. Sel. Evol.* 56:35. doi:10.1186/s12711-024-00892-9.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci* 97:3943–3952.

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202:401–409.

Vandenplas, J., J. ten Napel, S.N. Darbaghshahi, R. Evans, M.P.L. Calus, R. Veerkamp, A. Cromie, E.A. Mäntysaari, and I. Strandén. 2023. Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet. Sel. Evol.* 55:37. doi:10.1186/s12711-023-00808-z.