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

Andres Legarra\(^1,2\) (ad honorem), \texttt{andres.legarra@uscdcb.com} Matias Bermann\(^2\), Paul M VanRaden\(^3\),

Ezequiel L Nicolazzi\(^1\), Rodrigo R Mota\(^1\), Joe M Tabet\(^2\), Daniela Lourenco\(^2\), Ignacy Misztal\(^2\)

1 Council on Dairy Cattle Breeding, Bowie, MD

2 University of Georgia, Athens GA

3 AGIL, United States Department of Agriculture, Beltsville, MD
CDCB has

- 8M genotyped animals imputed at ~79K
- 100M animals in pedigree
- 30 “normal” (yield, health, calving ease...) traits + 20 “type” traits
- 8K to 50M animals in data depending on the trait
- 6 breeds highly unbalanced, crosses, all-breed evaluation
- We receive pedigree and genotypes from all over the world

Here I present choices to test & run ssGBLUP for CDCB
Trimming pedigree

• In BLUP, we only need animals in data and their ancestors
• a Python program trimming pedigree takes ~10 minutes
• in fertility data: reduction from ~100M to ~60M pedigree
Metafounder covariance for missing pedigree

• 5%-10% missing pedigree
• We used ~400 metafounders based on base allele frequencies (from imputation run) and increase of inbreeding (see recent GSE paper)
• a parallel analysis by Joe Tabet tried J-factors giving slightly more bias and noisier UPG solutions
Trimming genotypes in ssGBLUP

• for sure keep genotypes of animals that either “have records” or “have progeny with records”
• if not the case, do we keep them?
  • if both parents genotyped, the genotype provides $\emptyset$ information
  • if one (or both) parent(s) is NOT genotyped, the genotype improves a bit the $H$-relationship of parent(s)
  • we consider that this improvement is negligible, so we don’t keep them
• We therefore keep genotypes that either “have records” or “have progeny with records”:
  • reduction from 8M genotypes to 2M “useful” genotypes
• the US reference population now consists in 1.7M cows with records and >30K bulls with phenotyped offspring
Approaches for Single Step GBLUP

- SNP-based (Legarra and Ducrocq 2012, Liu et al., 2014, Fernando et al. 2016)
- No free lunch
- They usually involve either approximations, or some kind of blending, or complex programming
Approaches for Single Step GBLUP

• Why we like APY
• relationship-based
  • incidence matrices of effects are sparse
  • G_APY reduces *enormously* the size of genomic data to handle
  • fast convergence of iterative solvers for MME, good condition number
  • you can use “regular” double precision “multipliers” (Lapack, MKL, etc)
  • flexible: fractional genotypes, MIR readings, -omics …
Approaches for Single Step GBLUP

• Why we don’t like APY

• need to choose an informative, “repeatable” core
  • “random” is not a realistic or practical option for dairy
  • “proven bulls” is not any more a realistic option
  • no realistic way of doing matrix computations (e.g. PCA, Pocrnic et al. 2022) in 2M genotypes
Choice of Core

• We want a “democratic” core population
• For each breed you have a size of core
• Select “proven” genotyped bulls
  • Holstein: >500 daughters with records
  • other breeds >100 daughters with records
• select a sample of genotyped reference cows
  • tag cows with records
  • select 1 cow every n based on ID until fill in the available spots
<table>
<thead>
<tr>
<th>Breed</th>
<th># of Genotypes</th>
<th>Breed</th>
<th># of Sires+ Cows in Core</th>
<th>Core 98% eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ayrshire</td>
<td>1,608</td>
<td>AY</td>
<td>311 + 1,175 (all animals)</td>
<td></td>
</tr>
<tr>
<td>Brown Swiss</td>
<td>9,560</td>
<td>BS</td>
<td>611 + 4,313</td>
<td>5K</td>
</tr>
<tr>
<td>Guernsey</td>
<td>3,561</td>
<td>GU</td>
<td>219 + 3,258 (all animals)</td>
<td></td>
</tr>
<tr>
<td>Holstein</td>
<td>1,669,795</td>
<td>HO</td>
<td>6,890 + 8,113</td>
<td>15K</td>
</tr>
<tr>
<td>Jersey</td>
<td>300,976</td>
<td>JE</td>
<td>3,186 + 11,883</td>
<td>15K</td>
</tr>
<tr>
<td>Crossbreds</td>
<td>56,528</td>
<td>XX</td>
<td>141 + 4,616</td>
<td>5K</td>
</tr>
</tbody>
</table>

- Core: ~45K animals
- Non-core: 2M animals
- Reference population is the sum of core and non-core [and does include crossbreds]
Preparation of APY

• Once we have the flags core–noncore and the genotypes

• Build by chunks the $G_{\text{APY}}$ matrix (all by MKL Lapack)
  • Biggest chunk: $G_{\text{noncore,core}}$ and its reciprocal in $G_{\text{APY}}^{-1}$ of size 2M x 45K
  • $G_{\text{APY}}^{-1}$ stored double precision: $\approx 720$ Gigabytes
Memory mapping

• use “memory mapping” `mmap()` to handle $G_{APY}^{-1}$

• A memory-mapped file is a segment of virtual memory[^1] that has been assigned a direct byte-for-byte correlation with some portion of a file [...] this correlation between the file and the memory space permits applications to treat the mapped portion as if it were primary memory.

• 720 Gb RAM become 720 Gb disk

• modern alternative to “read from file and compute” iteration-on-data
Running of APY

• PreGSf90: Set up $G_{APY}^{-1}$ (with blending of [5% or 10%] $A_{T22}$).
  • RAM $\approx$ 720 Gigabytes [not using mmap()]

• Blup90iod3 (PCG iteration on data)
  • uses “memory mapping” mmap() to handle $G_{APY}^{-1}$
  • As a result, only 120 Gb (non-genomic parts, including the 4 x 60M animals GEBVs... ) are needed for the iteration

• accf90GS2 for reliabilities (Bermann et al 2022a) also uses mmap()

• backsolving SNP solutions only needs “core” animals (Bermann et al. 2022b)
Rough timings and memory

• Previous editings – perhaps 3h – may be improved
• 4 fertility traits, 50M records, 60M in pedigree, ~2M animals genotyped, ~500M equations

• using 16 threads
• Prep of $G_{APY}^{-1}$: 16h, 720 Gb RAM
• ssGBLUP itself: 22h, 120 Gb RAM, and 476 rounds of PCG
• Genomic reliabilities (include blending): ~8h per trait, 120 Gb RAM
• Backsolving for SNP solutions: negligible
• similar numbers as Cesarani et al. 2022
It works (Joe Tabet in prep.)

Slopes b1 within Interbull limits

Bias: ssGBLUP < BLUPMetafounders < (UPG + Jfactors)

LR correlations slightly better than (not shown here) 2-step at CDCB evaluations
Improvement?

• Maybe we don’t need it
  • we need to benchmark and play with number of threads
  • with some optimization in the pipeline and threads, time: ~2 days

• Anyway, some long-term perspectives just in case
  • $G_{APY}^{-1}$ fully run in mmap()
  • or, $G_{APY}^{-1}$ can be “updated” from previous runs
  • Reliability approaches can be optimized
Acknowledgments

- Participating dairy producers for supplying data
- DHI organizations and DRPCs for processing and relaying the information to the Council on Dairy Cattle Breeding (CDCB)
- Purebred breed associations for providing pedigree data

- Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by CDCB
- CDCB is an equal opportunity provider and employer

Picture L Chaumonnot