Estimation of across-breed metafounder parameters in all-breed US dairy cattle.

Andres Legarra[1,2], Matias Bermann[2]

[1]Council on Dairy Cattle Breeding

[2]University of Georgia

Interbull Open meeting, Verona, June 1st 2026

Metafounders: intro

- Use (sometimes misuse) of Unknown Parent Groups to model breeds and missing pedigree
- With single-step: J-factors (Stranden et al 2023) *vs*. Metafounders (Legarra et al 2015)
- J-factors assume that groups within breeds are “as different” than across breeds
- Metafounders give good results and they result in a more elegant and flexible theory
  - Legarra and VanRaden, 2023, Macedo et al (2021, 2022…) Wicky et al (2022, 2023) Tabet et al (2023) Kudinov (2021, 2022)…

CDCB™
COUNCIL ON DAIRY CATTLE BREEDING

Metafounders: intro

- Metafounders represent “pools” of founder individuals
- Unlike UPGs, Metafounders have self-relationships and across-relationships contained in a matrix $\boldsymbol{\Gamma}$.
- For instance, $\boldsymbol{\Gamma}\begin{pmatrix} Holstein \\ Jersey \end{pmatrix} = \begin{pmatrix} \gamma_{HOHO} & \gamma_{HOJE} \\ \gamma_{JEHO} & \gamma_{JEJE} \end{pmatrix}$
- **A** is built from $\boldsymbol{\Gamma}$ following tabular rules
- If done properly, there is automatic compatibility between **G** and **A**
- Where do we get $\boldsymbol{\Gamma}$ from? Markers and pedigree

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

State of the art for Metafounders $\boldsymbol{\Gamma}$

- Breeds + “crosses” + missing pedigrees 🤔
- Combine two ideas (Wicki et al 2023, Legarra et al 2024)
  - Estimate “baseline $\boldsymbol{\Gamma}$” across breeds using ML or a similar method:

$$\boldsymbol{\Gamma} = \frac{2}{k}(2\mathbf{P} - \mathbf{1}\mathbf{1}')(2\mathbf{P} - \mathbf{1}\mathbf{1}')'$$

we want p’s from the beginning of pedigree

  - Then expand within breed across years using “increase of relationships” based on year of birth

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Objective

This work compares methods to estimate "baseline $\boldsymbol{\Gamma}$" across the genotyped official breeds in US dairy: Ayrshire, Brown Swiss, Guernsey, Holstein and Jersey

Using imputation (findhap.f90)

- Imputation: use allele frequencies from old maternal haplotypes, then $\boldsymbol{\Gamma} = \frac{2}{k}\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)'$

Using the brand, new, efficient gammaf90 ☺

- GLS: (Garcia-Baccino et al. 2017) estimate allele frequencies using Gengler's (2007) method (BLUP gene content) or (equivalently)

$$\widehat{\mathbf{2p}}_i - \mathbf{1} = \left(\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{z}_i$$

$$\boldsymbol{\Gamma} = \frac{2}{k}\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)'$$

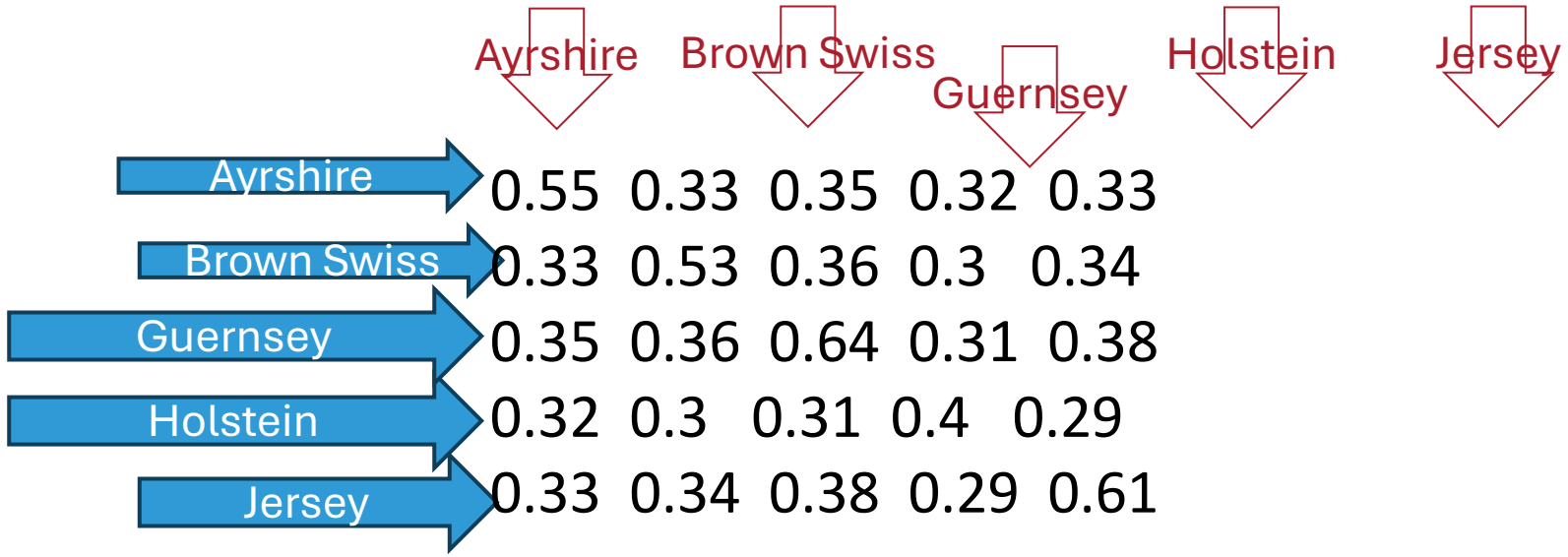
- Pseudo-EM: try to maximize likelihood of observed genotypes $\boldsymbol{z}_i$ {-1,0,1} , assuming normality and

$$Var\begin{pmatrix}\mathbf{2p}_i - \mathbf{1} \\ \boldsymbol{z}_i\end{pmatrix} = 0.5\begin{pmatrix}\boldsymbol{\Gamma} & \mathbf{A}_{\Gamma(mf,2)} \\ \mathbf{A}_{\Gamma(2,mf)} & \mathbf{A}_{\Gamma(2,2)}\end{pmatrix}$$

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

example baseline Γ for all breeds

using imputation-based ~1990 allele frequencies from the 2023 Dec run



Pseudo-EM, details

We start with an initial value $\boldsymbol{\Gamma} = 0.1\mathbf{I}$. At iteration t:

- Build (parts of) $\mathbf{A}_{\Gamma(t)}$ and obtain allele frequencies using

$$\widehat{\boldsymbol{2p}}_i - \mathbf{1} = \mathbf{A}_{\Gamma(\mathrm{t})(2,mf)}\mathbf{A}^{-1}_{\Gamma(\mathrm{t})22}\mathbf{z}_i$$

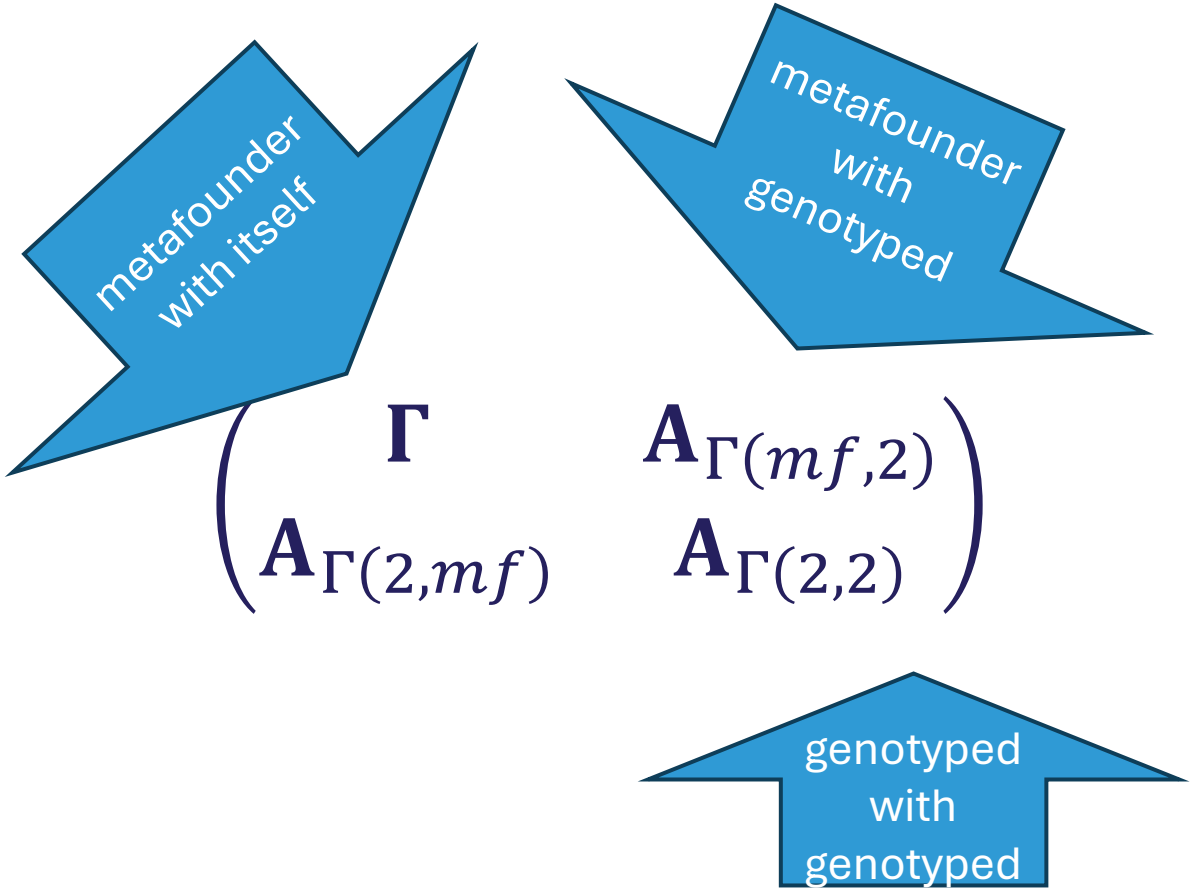
- Update the estimate $\widehat{\boldsymbol{\Gamma}}_{(t)}$ using

$$\widehat{\boldsymbol{\Gamma}}_{(t)} = \widehat{\boldsymbol{\Gamma}}_{(t-1)} - \mathbf{A}_{\Gamma(mf,2)(t)}\mathbf{A}^{-1}_{\Gamma 22(t)}\mathbf{A}_{\Gamma(t)\left(2\mathbf{A}_{\Gamma(t)(2.mf)}\right)} + \frac{2}{k}\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)'$$

PEV/PEC

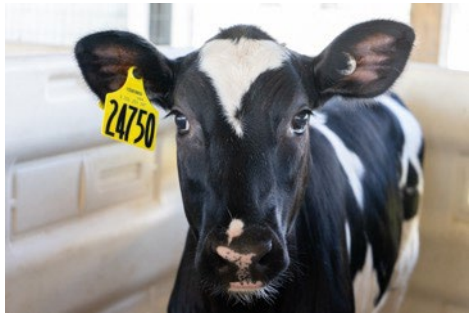
Crossproduct of estimated allele frequencies

Note, the “Pseudo-” because it ignores part of the likelihood – it might not find the optimum


metafounder with itself
metafounder with genotyped
$\boldsymbol{\Gamma}$
$\mathbf{A}_{\Gamma(mf,2)}$
$\mathbf{A}_{\Gamma(2,mf)}$
$\mathbf{A}_{\Gamma(2,2)}$
genotyped with genotyped

Data





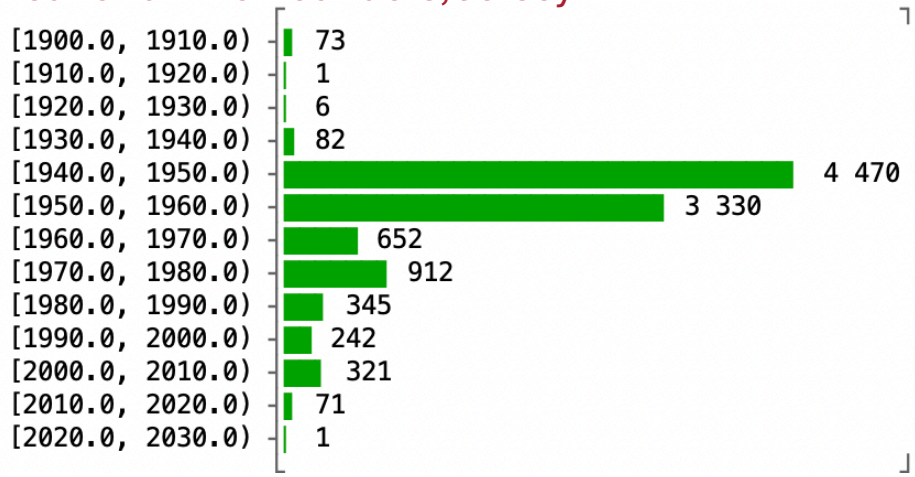
Data

- There are >10M genotyped animals at CDCB
- For GLS, Pseudo-EM we used a subset from the “gestation length” evaluation subpedigree
- ~16K genotypes, mostly sires and dams-of-sires
- No crossbreds
- 380K pedigree !!
- Highly unbalanced genotyping
- “Mode” of year of birth of “founders of pedigree” ~1950

Repartition of genotyped animals per decade and breed

Year	AY	BS	GU	HO	JE
1950	2	3	0	7	4
1960	8	10	5	20	14
1970	17	25	8	60	22
1980	37	47	24	232	112
1990	92	193	59	854	559
2000	87	128	79	1750	624
2010	101	257	55	9649	2008
2020	6	18	7	920	198

Year of birth of founders, Jersey



Results: general

- Pseudo-EM increased logL steadily until an asymptote
- Pseudo-EM and Imputation agreed well (cor elements of $\boldsymbol{\Gamma}$ 0.96, maximum difference <0.1)
- GLS overestimates self-values of $\boldsymbol{\Gamma}$ for small breeds (e.g. AY)
- In other words, with GLS small breeds look too inbred
  - errors in $\widehat{2\boldsymbol{p}}$ become squared

Results: across breeds

- With GLS, breeds look MUCH more different than they are, Fst's are too high
- Pseudo-EM looks OK
- Typical values of Wright's Fst differentiation index are 0.10 - 0.20 (Gibbs et al 2009)

Table 2. Estimated Fst values across breeds.

	AY	BS	GU	HO	JE
AY		0.52	0.38	0.39	0.47
BS	0.17		0.32	0.34	0.41
GU	0.15	0.15		0.19	0.24
HO	0.12	0.14	0.11		0.27
JE	0.18	0.17	0.14	0.13	

Upper triangular: GLS.

Lower triangular: Pseudo-EM

Results: convergence

- Pseudo-EM is slow in convergence (1000 iterations) but it was smooth
- Some elements of $\mathbf{\Gamma}$ converged very quickly but others did not

Results: allele frequencies

- Spread of estimates of allele frequencies were *very* different
- GLS can give $\hat{p}$ out of the [0-1] bounds
  - They were restricted to [0-1] for estimation
- The problem exacerbates in small breeds
- Pseudo-EM induces much stronger shrinkage than GLS

Results: allele frequencies

- Imputation and Pseudo-EM frequencies agree very well
- GLS frequencies do not

Brown Swiss

Discussion (1)

- It's hard to estimate well 69200 *p*'s for 5 breeds = 350,000 unknowns, specially back in 1950!
- Allele frequencies from imputation were fine to get $\boldsymbol{\Gamma}$.
  - Maternal haplotypes from old animals
- Pseudo-EM gave very good results too
  - It is the default in gammaf90
  - It compensates automatically for lack of information
- GLS did not give good results, with obvious biases
  - Except for large breeds
- Methods and gammaf90 work with crossbreds too (but we didn't include them)

Discussion (2)

- Is all this relevant for genomic prediction?
  - I think having a good $\boldsymbol{\Gamma}$ is safer… but I can't really say…
  - Why do things wrong if you can do them right ?
  - It gives a recipe for complicated all-breed genomic predictions like US ones
- Is this relevant for diversity?
  - There' still some bias because SNP chips are tailored to some breeds
  - Diversity measures (heterozygosity, Fst) from Imputation or pseudo-EM were fine. From GLS, they were misleading

Acknowledgments

- The contribution of dairy producers who supplied data through their participation in the Dairy Herd Improvement program and the Dairy Records Processing Centers is acknowledged.