Efficient computation of base generation allele frequencies

11 February; Interbull meeting, Auckland, New Zealand

Michael Aldridge, Jeremie Vandenplas & **Mario Calus**

WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Allele frequencies in genomic prediction

- Genomic prediction requires allele frequencies (AF)
- Commonly, AF are current data averages
- Theoretically, AF should be computed for the base generation

Base generation AF

Base generation = base generation in pedigree!

Base generation AF required for calculation of:

- Genomic relationships in (single-step) GBLUP
- Model-based reliabilities for multi-step genomic evaluations
- Computation of relationships among metafounders[1]

[1]Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. *Genetics*. 200:455-468.

Objective

Compare accuracy and efficiency

of different methods to compute

base generation allele frequencies

Methods – overview

- AF: $p = \frac{1}{2}\hat{\mu}$

Method	Mean is estimated:
All	Across all genotypes
Oldest	Across oldest generation genotyped
BLUP	In BLUP model
GLS	General Least Squares (GLS)

Methods - BLUP

- BLUP model; y = genotype (0,1,2)
- h^2=0.99; allowing some genotyping error
- Univariate; or multivariate with zero genetic correlations
- Implemented using MiXBLUP

McPeek, M. S., X. D. Wu, and C. Ober. 2004. Biometrics. 60:359-367.
Gengler, N., P. Mayeres, and M. Szydlowski. 2007. Animal. 1:21-28.

Methods – GLS (dense / sparse)

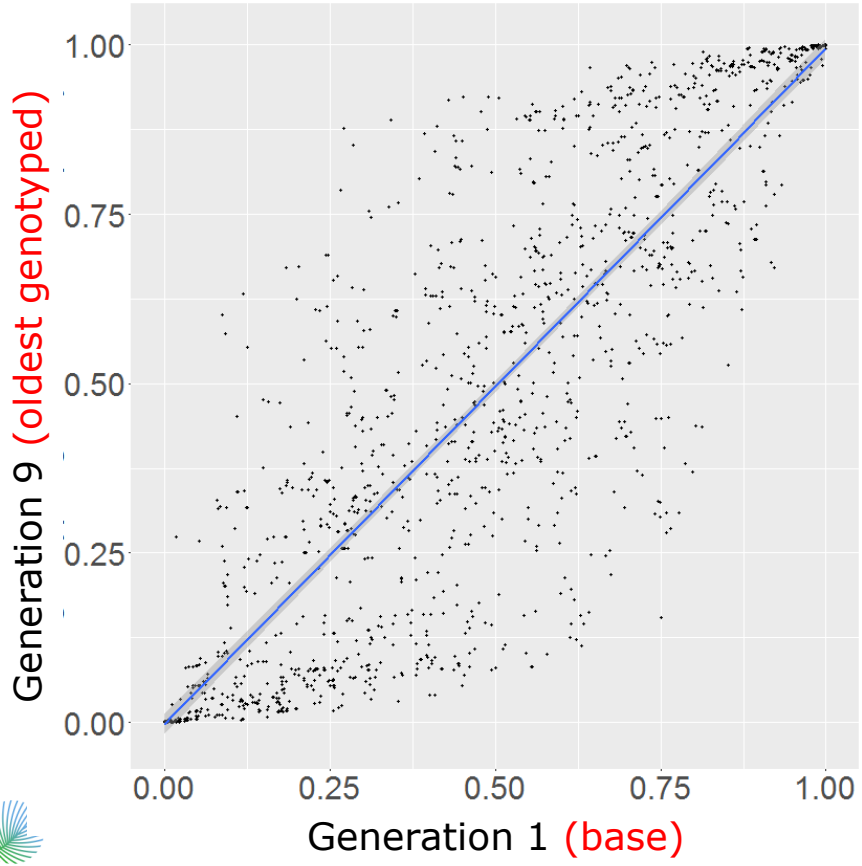
- GLS: $\hat{\mu}_i = (\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{Z}_i$
- Dense: Compute and invert $\mathbf{A}_{22}$ — Calc_grm
- Sparse: $\mathbf{A}_{22}^{-1}\mathbf{1} = \left(\mathbf{A}^{22} - \mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}\right)\mathbf{1}$ — Own program / Intel MKL-PARDISO

McPeek, M. S., X. D. Wu, and C. Ober. 2004. Biometrics. 60:359-367.
Garcia-Baccino, C.A., Legarra, A., Christensen, O.F., Misztal, I., Pocrnic, I., Vitezica, Z.G., and Cantet, R.J. 2017. Genet. Sel. Evol. 49, 34.

Data (simulation)

- Holstein-like population
- Generations 9 to 12 (after base) fully genotyped
- 325,266 animals in pedigree; 100,078 genotyped
- 1670 SNPs (providing replication)
- Selection: None or Strong

Change in AF across generations (with selection)



WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

Results - accuracy

Method	Without selection	With selection
All	0.99 ± 0.01	0.87 ± 0.01
Oldest	0.99 ± 0.01	0.88 ± 0.01
BLUP	0.99 ± 0.01	0.96 ± 0.01
GLS_dense	0.99 ± 0.01	0.97 ± 0.01
GLS_sparse	0.99 ± 0.01	0.97 ± 0.01

Results - efficiency

Method	Process time	RAM
All	0-00:03:44	7.8 GB
Oldest	0-00:01:19	1.6 GB
BLUP (60 SNPs)	0-13:42:17	49.0 GB
GLS_dense	50-20:12:16	165.9 GB
GLS_sparse	0-00:01:28	2.6 GB

=> Efficiency of GLS_sparse is very competitive!

Discussion

- Few GLS_sparse estimates outside 0-1 range:
  - Only for very low MAF <0.001
  - Swapping allele code solved most of those

- Estimates were not affected when having:
  - 2% genotyping errors
  - 25% of sires unknown

Conclusions

- Base generation AF required for:
  - Genomic relationships in (single-step) GBLUP
  - Model-based reliabilities for multi-step genomic evaluations
  - Computation of relationships among metafounders

- GLS_sparse estimator recommended
  - Accurate & very efficient

Acknowledgements