Single-step evaluation for calving traits with 1.5 million genotypes: APY and ssGTBLUP approaches

I. Strandén[1], R. Evans[2] & E.A. Mäntysaari[1]

[1]*Natural Resources Institute Finland – Luke*

[2]*Irish Cattle Breeding Federation, Cork, Ireland*

Ismo.Stranden@Luke.fi

Connected studies use the same data:

From One to Many: Re-Defining Calving Evaluations to Cater for Divergent Cow Types

Ross Evans, A.Cromie, S.Ring, T.Pabiou
Irish Cattle Breeding Federation, Highfield House, Bandon, Cork, Ireland

1) Poster Session 44, p. 468 (title changed!)
2) **This presentation**, p. 212
3) Theatre Session 12, p. 212

Single-step evaluation for calving traits with 1.5 million genotypes: SNP-based approaches

J. Vandenplas, R. Veerkamp, R. Evans, M.P.L. Calus & J. ten Napel

Keywords:
Beef cattle
Multiple breeds
Calving difficulty
Genomic data
Single-step
High performance computing

Multiple trait model: direct and maternal genetic effects

Heritability & Genetic Correlations

Heritability Direct/Maternal	Dairy Heifer	Dairy Cow	Beef Heifer	Beef Cow	Body Size	Body Weight
Dairy Heifer	0.16 / 0.04	0.76	0.39	0.57	0.34	0.61
Dairy Cow	0.91	0.08 / 0.02	0.75	0.73	0.82	0.67
Beef Heifer	0.80	0.78	0.17 / 0.09	0.97	0.66	0.56
Beef Cow	0.62	0.59	0.94	0.15 / 0.08	0.81	0.38
Body size	0.82	0.74	0.88	0.85	0.24 / 0.05	0.54
Body weight	0.63	0.64	0.64	0.62	0.52	0.41 / 0.09

Correlations: direct below diagonal, maternal above diagonal

Small heritabilities for the calving difficulty traits, larger for the correlated traits.

Phenotypes (data until 2015)

Trait	N	Mean	SD	Min	Max
Dairy Heifer	604,323	1.43	0.68	1	4
Dairy Cow	1,914,276	1.30	0.58	1	4
Beef Heifer	158,578	1.64	0.84	1	4
Beef Cow	512,357	1.45	0.72	1	4
Body size	1,814	3.11	0.80	2	5
Body weight	69,290	41.44	7.49	20	100

Number of pedigree animals: 10.36 million
Number of data records: 3,221,888

Number of genotyped (used in the analysis): 1,512,383
Number of markers: 50,855

Model information

- Direct and maternal genetic effects for all traits
  - Genetic groups (breed fractions) by regression
- Models:
  - Animal model (AM), no genomics
  - ssGTBLUP:
    - 98% of variation, based on eigenanalysis
      - giving 31,443 components out of 50,855
        == number of rows in **T** matrix
    - 99% of variation: 35,356 components

Number of unknowns:
129,349,562

T matrix in memory (#components by 1.5M matrix)

Model information

- Direct and maternal genetic effects for all traits
  - Genetic groups (breed fractions) by regression
- Models:
  - Animal model (AM), no genomics
  - ssGTBLUP:
    - 98% of variation, 31,443 components
    - 99% of variation, 35,356 components
  - ssGBLUP-APY:
    - APY31K, 31,443 random core animals
    - APY35K, 35,356 random core animals
    - Inverse $\mathbf{G}_{APY}$ matrix in memory

Number of unknowns: 129,349,562

Note: 10 processors were utilized in computations

Genomic relationship matrix G

- VanRaden method I: centered marker matrix **Z**
  - Base population allele frequencies (by GLS method)
- **G** matrix scaled by tr($\mathbf{A}_{22}$)/tr(**G**)
- Value 0.01 added to the diagonal elements
  ➔ **G** matrix non-singular

Note: In reality **G** matrix was never made:

- Making of **T** matrix does not need it
- APY approach was done by a memory efficient way

Making T/inv(G) for ssGTBLUP/APY

	Peak memory (GB)	Wall clock time (h)	Most time consuming
Te, 98%	374	10.3	**Z'Z**: 3.4h, eigenv. (**V**): 1.9, **VZ**: 2.7h
Te, 99%	382	10.3	**Z'Z**: 3.5h, eigenv. (**V**): 1.9, **VZ**: 2.4h
APY31K	558	7.7	$\mathbf{G}_{\mathbf{1.5M,31K}}$ make: 3.3h, inverse: 3.8h
APY35K	627	9.6	$\mathbf{G}_{\mathbf{1.5M,35K}}$ make: 4.2h, inverse: 4.8h

Notes:
- APY is memory efficient version where the matrix is done in parts
- **VZ** computations read marker matrix **Z** in 3 parts to save memory

Strandén et al., Single-step evaluations for 1.5M genotyped, EAAP 2019

Software

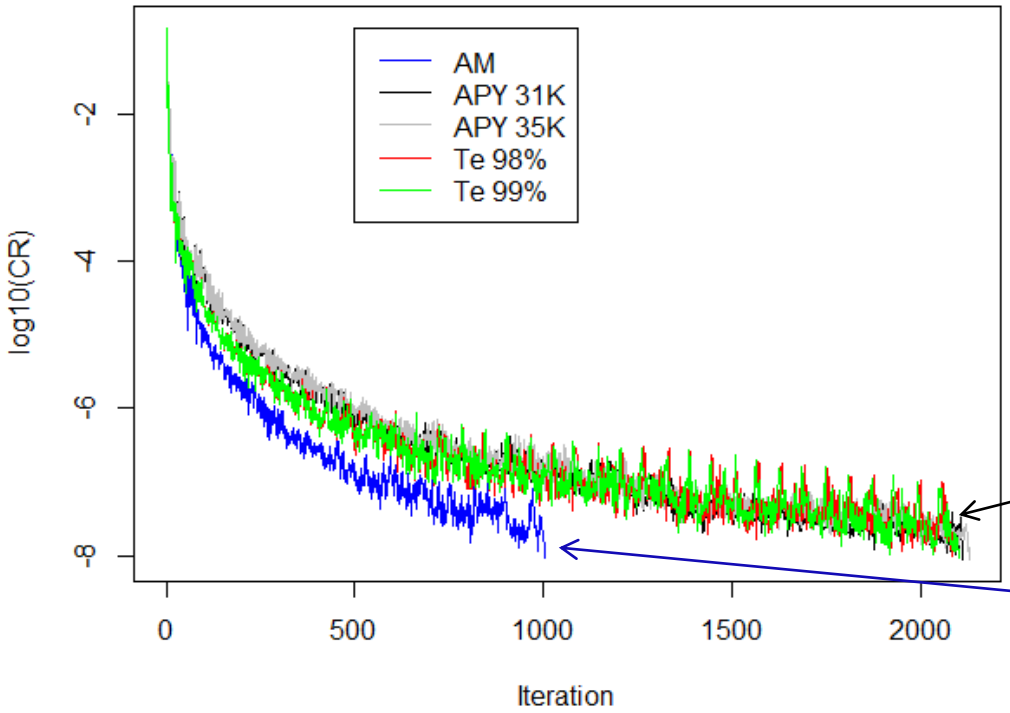
- PCG solver (MiX99) using iteration on data

$$\text{Convergence statistic: } \mathrm{CR} = \sqrt{\frac{(\mathbf{Cx}-\mathbf{b})'(\mathbf{Cx}-\mathbf{b})}{\mathbf{b}'\mathbf{b}}} < 10^{-8}$$

- $(\mathbf{A}_{22})^{-1}$ computations using sparse Cholesky
- Genomic matrices ($\mathbf{T}/\mathbf{G}^{-1}$) in memory to allow parallelization by BLAS MKL library subroutines
  - Matrix products
- Multi-threaded limited to 10 threads

Convergence: CR < 10^{-8}

Convergence statistic: $\mathrm{CR} = \sqrt{\frac{(\mathbf{Cx}-\mathbf{b})'(\mathbf{Cx}-\mathbf{b})}{\mathbf{b}'\mathbf{b}}}$



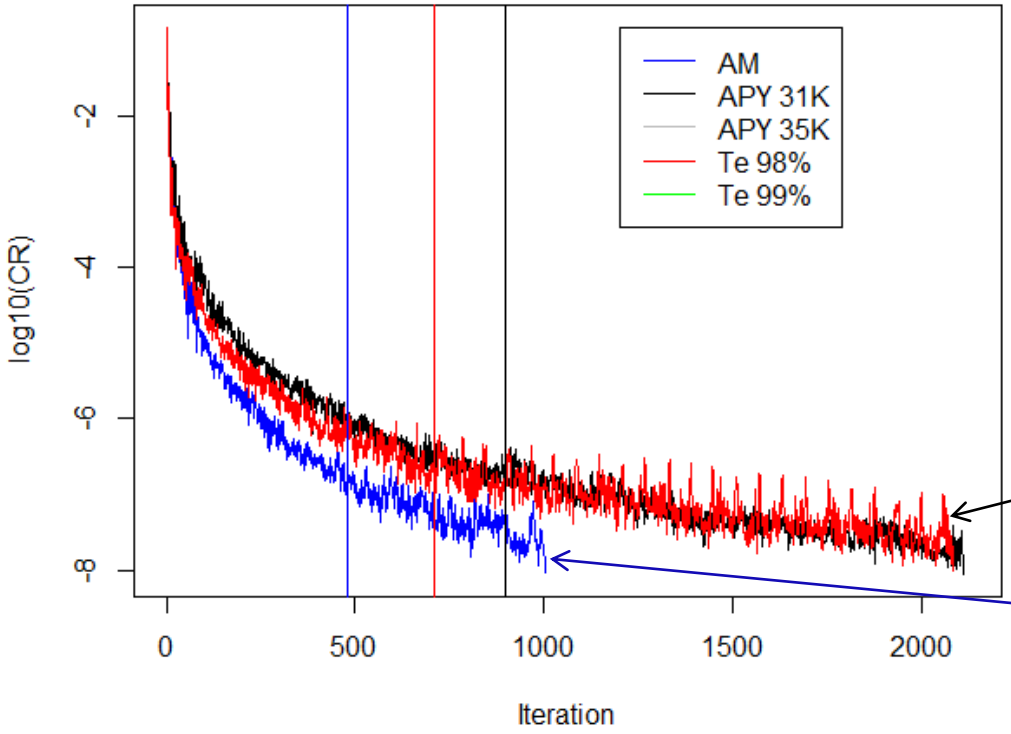
ssGBLUP models

Animal model, no genomic

Strandén et al., Single-step evaluations for 1.5M genotyped, EAAP 2019

Convergence: how many iterations is enough?

Convergence statistic: $\mathrm{CR} = \sqrt{\frac{(\mathbf{Cx}-\mathbf{b})'(\mathbf{Cx}-\mathbf{b})}{\mathbf{b}'\mathbf{b}}}$



ssGBLUP models

Animal model, no genomic

CR<10^{-7}, N iterations: 479 by AM, 897 by APY 31K, 710 by ssGTBLUP 98%

Correlation of genotyped animal solutions: 98% and 31K cases

	ssGTBLUP CR < 10^{-8} vs. 10^{-7}		ssGTBLUP vs. APY CR < 10^{-8}	
Trait	Animal	Maternal	Animal	Maternal
Dairy Heifer	0.982	0.965	0.992	0.992
Dairy Cow	0.978	0.960	0.993	0.990
Beef Heifer	0.953	0.960	0.987	0.988
Beef Cow	0.988	0.964	0.985	0.987
Body size	0.971	0.965	0.988	0.989
Body weight	0.965	0.943	0.990	0.991

Strandén et al., Single-step evaluations for 1.5M genotyped, EAAP 2019

Solver program performance (CR < 10^{-8})

Approach	Peak Mem (GB)	Time/iter (sec)	N iter	Solver time (h)
AM	4	9	1003	2.6
ssGTBLUP, 98%	366	51	2084	30.7
ssGTBLUP, 99%	410	58	2103	35.1
APY, 31K in core	366	55	2109	33.2
APY, 35K in core	410	60	2128	36.9

Similar memory need and computing time per iteration.

Caution: WALL CLOCK computing times are affected by other users.

Conclusions

- Single-step model increased preprocessing time (at most 10h), and solver time from AM 2.6h to ssGTBLUP >30h

- Animal model without genomic relationship matrix (AM) converged faster (half N iter.) than any of the ssGBLUP approaches
  - Need to improve the preconditioner
  - Currently: diagonal having trait (6 x 6) blocks by effect level
  - Convergence by CR needs to be quite tough (CR < 10^{-8})

Conclusions

- Correlations in GEBV between APY and ssGTBLUP were only >98.5% between 31K core and 98% eigenvalues but >99.1% between 35K core and 99% eigenvalues

- APY and ssGTBLUP had similar performance: APY had slightly lower preprocessing time than ssGTBLUP but needed a bit more solver time

Thank you!