Genomic Preselection and Future MACE

Pete Sullivan (Lactanet, Canada)

What is Genomic Pre-Selection (GPS)?

- GPS is when we choose only a subset of genotyped candidates for phenotyping.
- GPS alters the distributions of true BV for phenotyped individuals in our GE systems
- Distributions of true BV for GPS groups of individuals have shifted means and reduced variances relative to the full normal distribution for all candidates prior to selection
- GBLUP can account for GPS effects on true BV distributions, if the genotypes of all selection candidates are included
- GBLUP can therefore generate unbiased GEBV

NOTE: my "GBLUP" here can include Multi-step (G) and Single Step (H) systems, animal-based and SNP-based parameterizations: G-BLUP, H-BLUP, SNP-BLUP, …

GPS effects in PBLUP systems

- PBLUP systems do not include genotypes, but phenotypes are eventually recorded, and the phenotypes include expression of the GPS effects.
- Modified distributions can be estimated for GPS groups of individuals from phenotypes
- We have a problem, however, that PBLUP doesn't know if observed distributional changes (e.g. in elevated phenotypic means) were due to GPS of the sires, or due to other factors in the model, like herd environment effects, genetic value of the sire's mates, Mendelian sampling of the daughters, PA vs MS of the sire, etc.
- The EBV of a GPS sire, his mates and progeny from PBLUP are probably all biased if we do not, in some way, fully direct sire GPS effects into the sire's EBV, and away from these other individuals and environmental factors included in the model

NOTE: my PBLUP here refers to Pedigree-BLUP with no genotypes

So why use PBLUP in MACE?

- If instead of using PBLUP, we fed unbiased national GEBV into MACE, and then MACE into national GBLUP, we would repeatedly double-count the genomic information
- Although national EBV are biased they are also genomics-free, which allows the use of MACE proofs as input for national GEBV without double-counting the genotypes
- A Working Group was established in 2018 to work on solutions for reducing EBV bias while continuing to exclude individual genotype effects in a better future MACE service
  - First report from the future MACE WG, 2019 Interbull Meeting in Cincinnati, USA
  - Proposed model for Future MACE, 2022 Interbull Meeting in Montreal, Canada
  - Implementing a GPS-MACE service, 2023 Interbull Workshop in Rome, Italy

Key Reports and Activities

- JDS: National EBV are biased w/o genotypes used for GPS (Patry and Ducrocq, 2011)
- JDS: MACE proofs include the national EBV bias (Patry et al, 2013)
- Interbull workshop: Adapting MACE for GPS (Slovenia, 2017 Feb)

Interbull Technical Committee and Working Groups: tasked to quantify GPS effects and simulate GPS data to test future MACE approaches (Estonia, 2017 Aug)

- Interbull meeting: Modifying MACE for GPS (USA, 2019 Jun) ← Literature Review
- Interbull webinar: Genomic-free input for MACE (2021 Feb)
- Interbull meeting: Genetic regressions for GPS in MACE (Canada, 2022 May)
- Interbull workshop: Plans for implementing GPS-MACE (Italy, 2023 Feb)

Genomic-free input for MACE

Slide from 2021 Interbull Webinar

- Trade-off between no GPS-bias versus genomic-free
- "Genomic" preselection bias is mainly an early proof problem, which decreases with more daughters
- "Foreign-proof" preselection bias is an old, similar problem
- Future MACE working group seeks to reduce GPS bias with better (MS) model assumptions, focusing on both:
  - Improved MACE modeling + better MACE input data

Genomic-free EBV for MACE *(Interbull Webinar, Feb 11, 2021)*

Presentation Today

Discussion Today

Today's Presentations

1. Selection bias is generally not a big concern if all data used for selection can be properly included in a *Closed evaluation system* (I. Jibrila)
   - National Single-step without integration of foreign data
   - Breeding Company systems based on closed-line breeding
2. *Open system* data exchange/integration adds complexity (P. Sullivan)
   - Single-step with MACE integration for foreign sires
   - MACE with integration of national EBV without genotypes
   - GMACE, Intergenomics and SNP-MACE
3. *Software tools* and modeling approaches are available (I. Strandén)

Genomic preselection in single-step evaluation

Ibrahim Jibrila, Mario Calus, Gerben de Jong

Interbull Technical Workshop, 15/03/2023, Rome

WAGENINGEN UNIVERSITY & RESEARCH

INTERBULL

CRV Coöperatie

BETTER COWS > BETTER LIFE

Part 1: Impact of genomic preselection on accuracy and bias in subsequent single-step evaluation of preselected animals

Simulated breeding programme

- Single-trait breeding goal
- 15 recent generations with selection
- Pedigree: generations 0 to 15
- Genotypes: generations 13 to 15
- Phenotypes: generations 11 to 15

QMSim

Implementation of genomic preselection

Genomic preselection

- No preselection → Control scenario
- High preselection → 10% of males, 15% of females preselected
- Very high preselection → 5% of males, 12.5% of females preselected

Single-step evaluation used to preselect!

WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Subsequent evaluation

- Both PBLUP and ssGBLUP implemented
- $y_i = \mu + animal_i + e_i$
- $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & (0.9\mathbf{G} + 0.1\mathbf{A_{22}})^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$
- All information on preculled animals discarded

Measuring the impact of preselection on the subsequent evaluation

- Accuracy ($r_{TBV,GEBV}$)
- Level bias (mean TBV – mean GEBV)
- Dispersion bias ($b_{TBV,GEBV}$)

Summary of results from the Part 1

- Accuracy always reduced with preselection
- No bias with single-step, regardless of preselection scenario

Jibrila *et al. Genet Sel Evol* (2020) 52:42
https://doi.org/10.1186/s12711-020-00562-6

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Check for updates

Investigating the impact of preselection on subsequent single-step genomic BLUP evaluation of preselected animals

Ibrahim Jibrila*, Jan ten Napel, Jeremie Vandenplas, Roel F. Veerkamp and Mario P. L. Calus

WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Part 2: Information needed in subsequent single-step evaluations to prevent genomic preselection bias

Implementation of genomic preselection

- Same as in part1
- Exception: now only high genomic preselection scenario implemented

Nine scenarios based on sources and amounts of genomic information:

Preculled animals (G15)

Preselected animals (G15)

Culled sibs of parents of selection candidates (G14)

Parents of selection candidates (G14)

Grandparents of the selection candidates and their sibs (G13)

Four scenarios based on sources and amounts of phenotypic information:

Preselected animals (G15)

G14

G13

G12

G11

With both genotypes and phenotypes at preselection

With phenotypes but no genotypes at preselection

Summary of results from Part 2

- To prevent preselection bias in subsequent single-step evaluations, the following are needed:
  - Reference data used at preselection stage
  - Genotypes and of preselected animals
- Genotypes of preculled animals only needed if their parents are not genotyped!

Received: 12 August 2020 | Revised: 21 November 2020 | Accepted: 9 December 2020
DOI: 10.1111/jbg.12533

Check for updates

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics | WILEY

Avoiding preselection bias in subsequent single-step genomic BLUP evaluations of genomically preselected animals

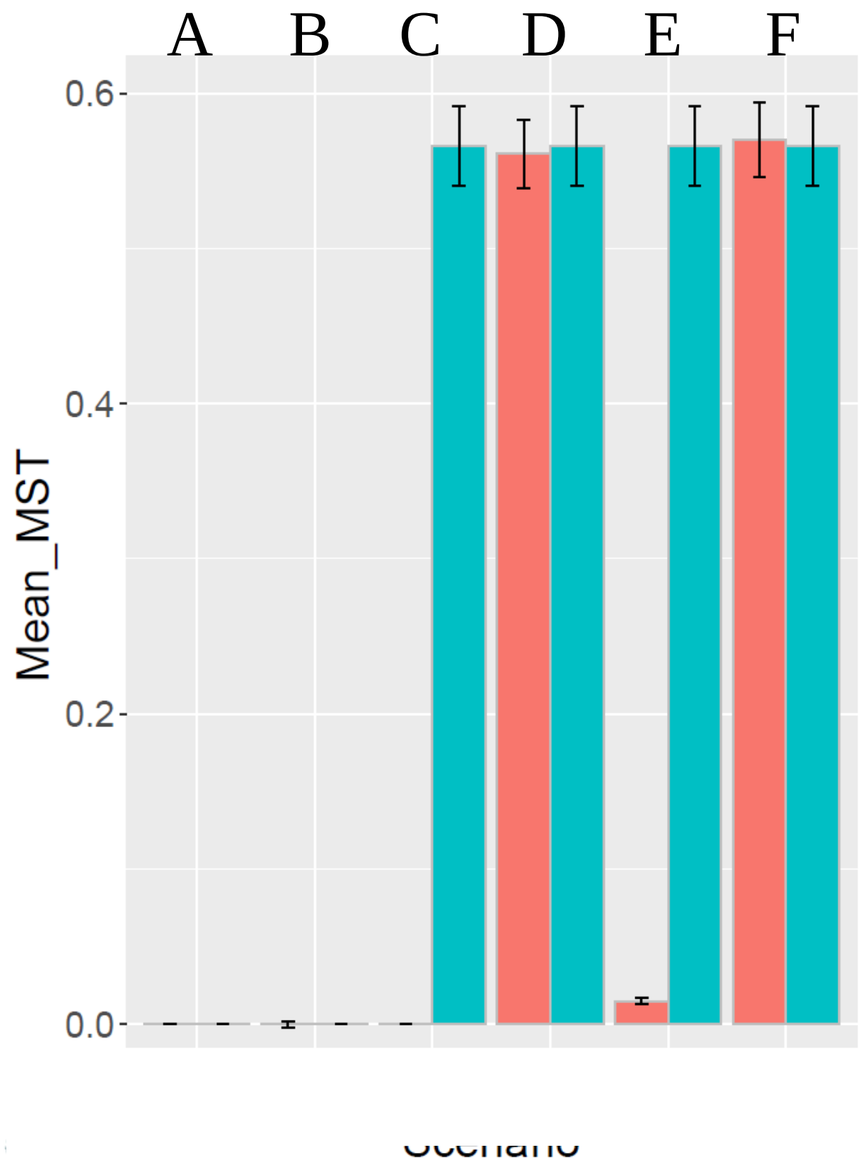
Ibrahim Jibrila | Jeremie Vandenplas | Jan ten Napel | Roel F. Veerkamp | Mario P. L. Calus

WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Part 3: Single-step prevents preselection bias in subsequent evaluation by correctly estimating Mendelian sampling terms of preselected animals

Averages of Mendelian sampling terms



A ➔ Subsequent pedigree eval., Ctrl scenario

B ➔ Subsequent single-step eval., Ctrl scenario

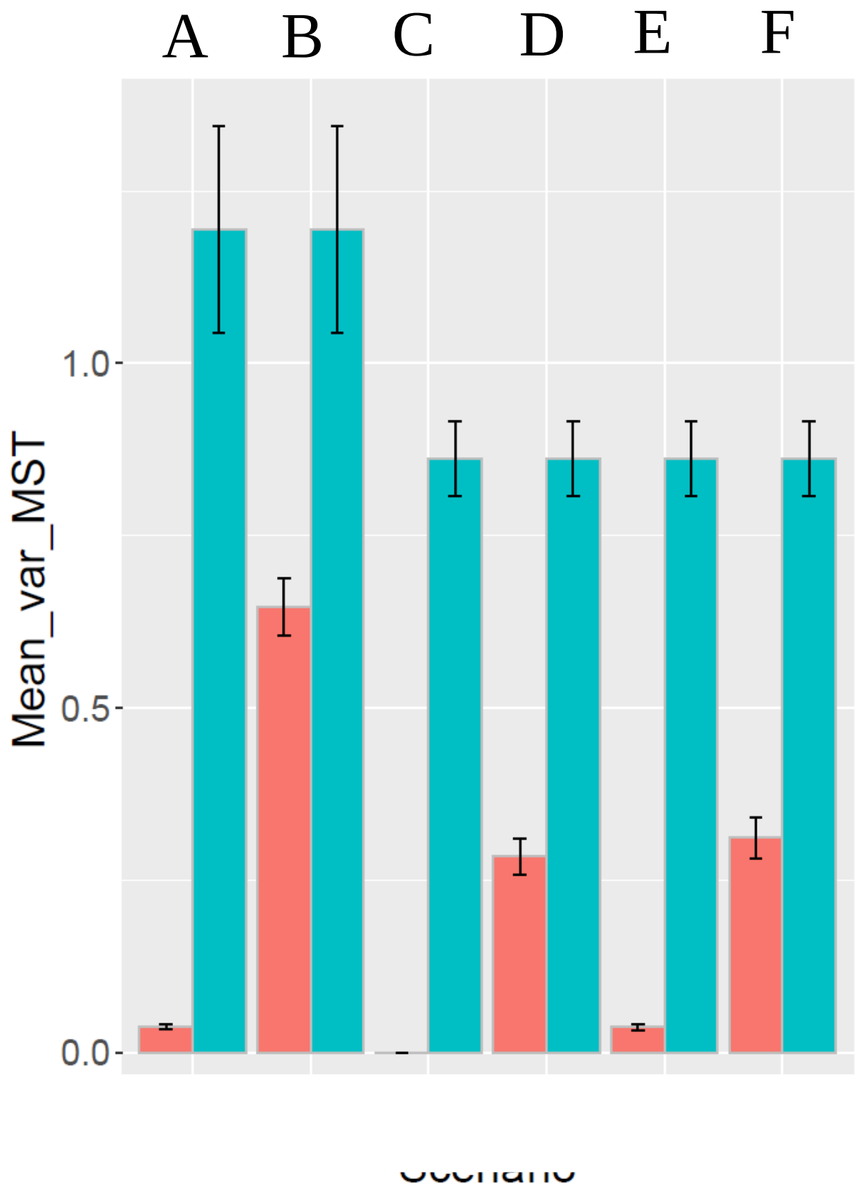
C ➔ Initial pedigree eval., GPS scenario

D ➔ Initial single-step eval., GPS scenario

E ➔ Subsequent pedigree eval., GPS scenario

F ➔ Subsequent single-step eval., GPS scenario

Variancesof Mendelian sampling terms



MST_type

Estimated

True

A ➔ Subsequent pedigree eval., Ctrl scenario

B ➔ Subsequent single-step eval., Ctrl scenario

C ➔ Initial pedigree eval., GPS scenario

D ➔ Initial single-step eval., GPS scenario

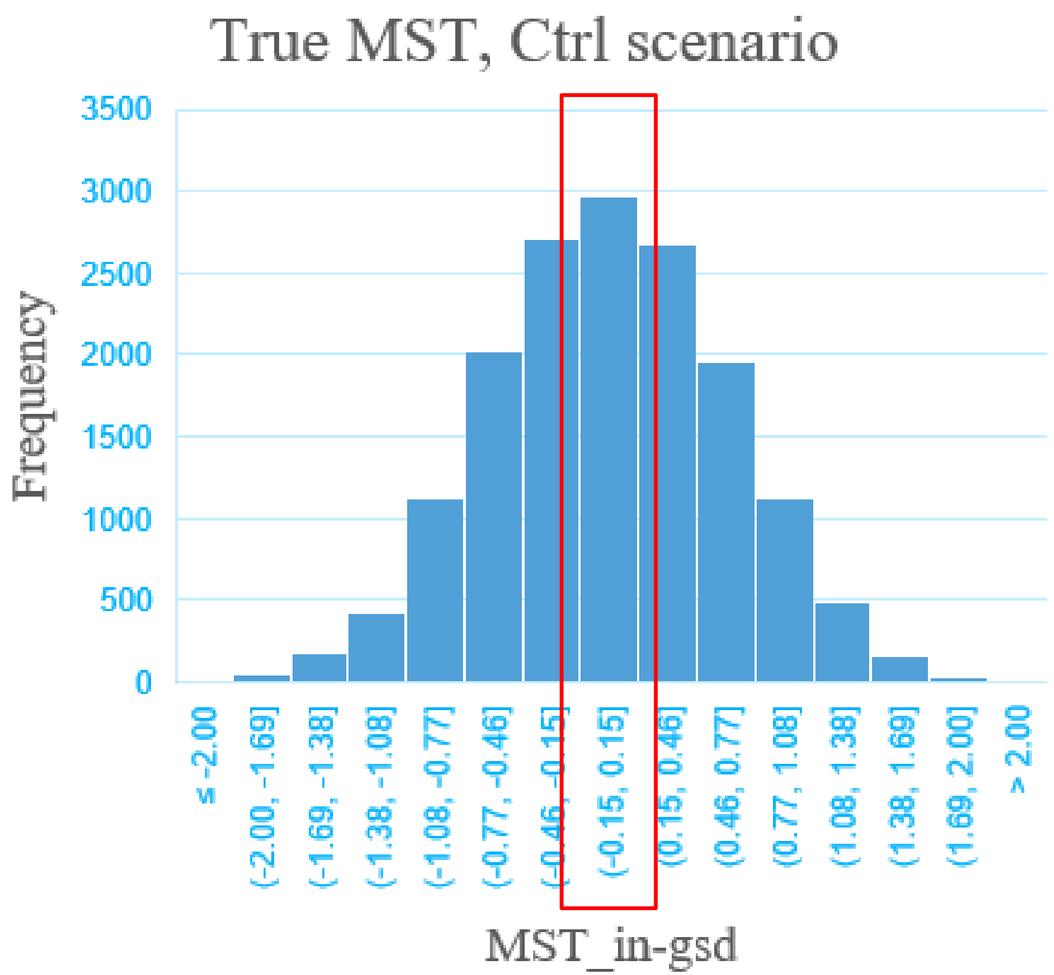
E ➔ Subsequent pedigree eval., GPS scenario

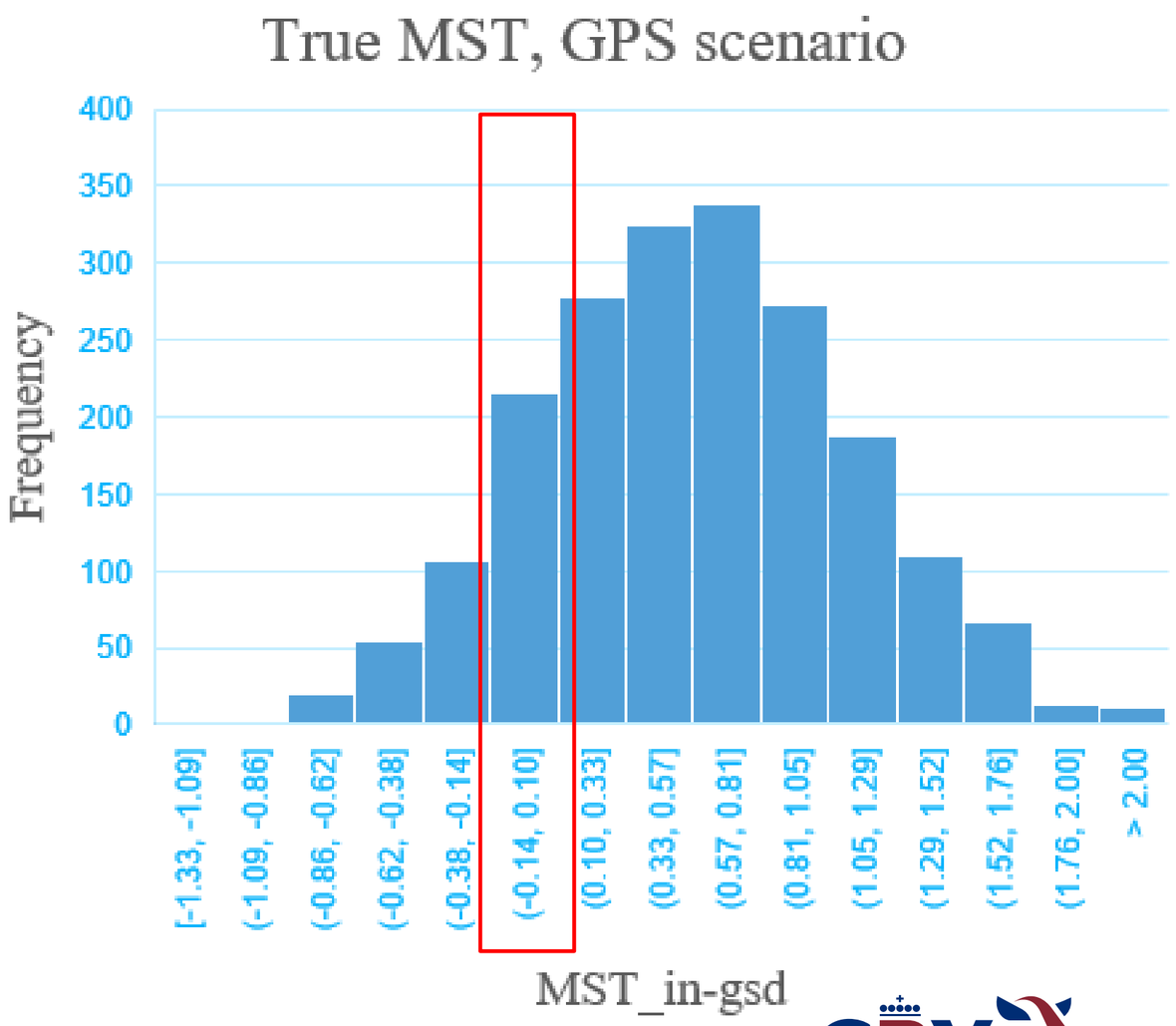
F ➔ Subsequent single-step eval., GPS scenario

CRV Coöperatie

BETTER COWS > BETTER LIFE

Distributions of Mendelian sampling tems

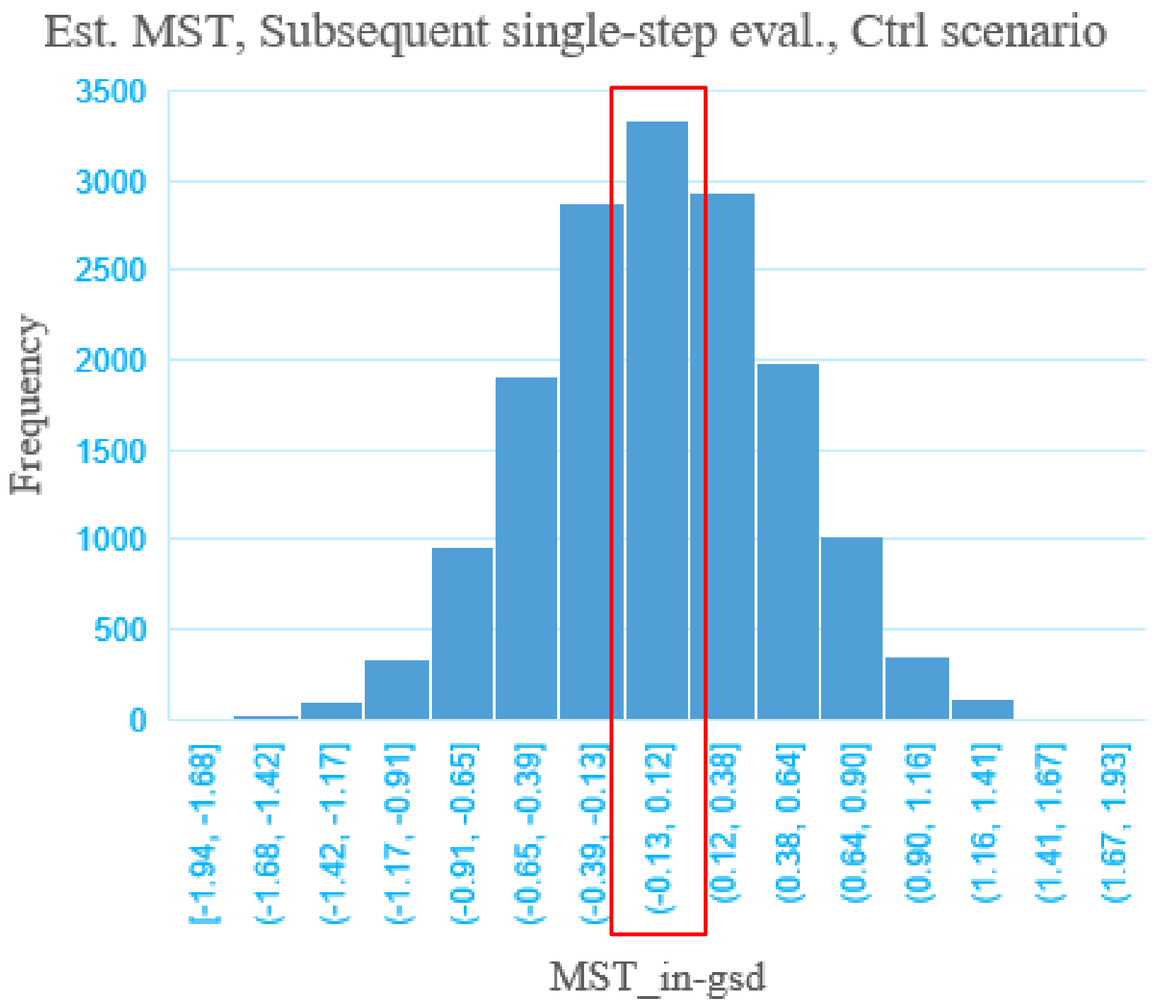


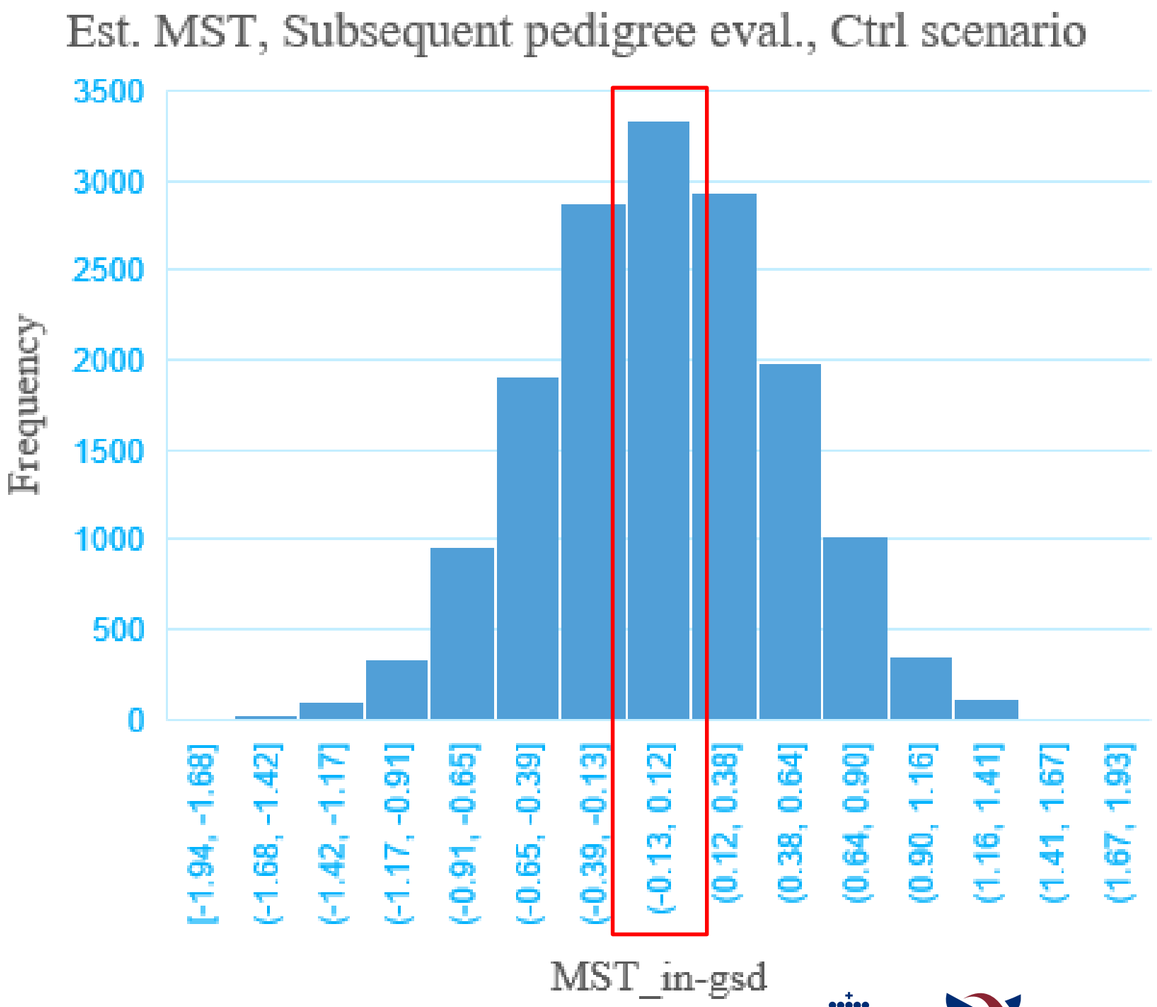


WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Distributions of Mendelian sampling tems

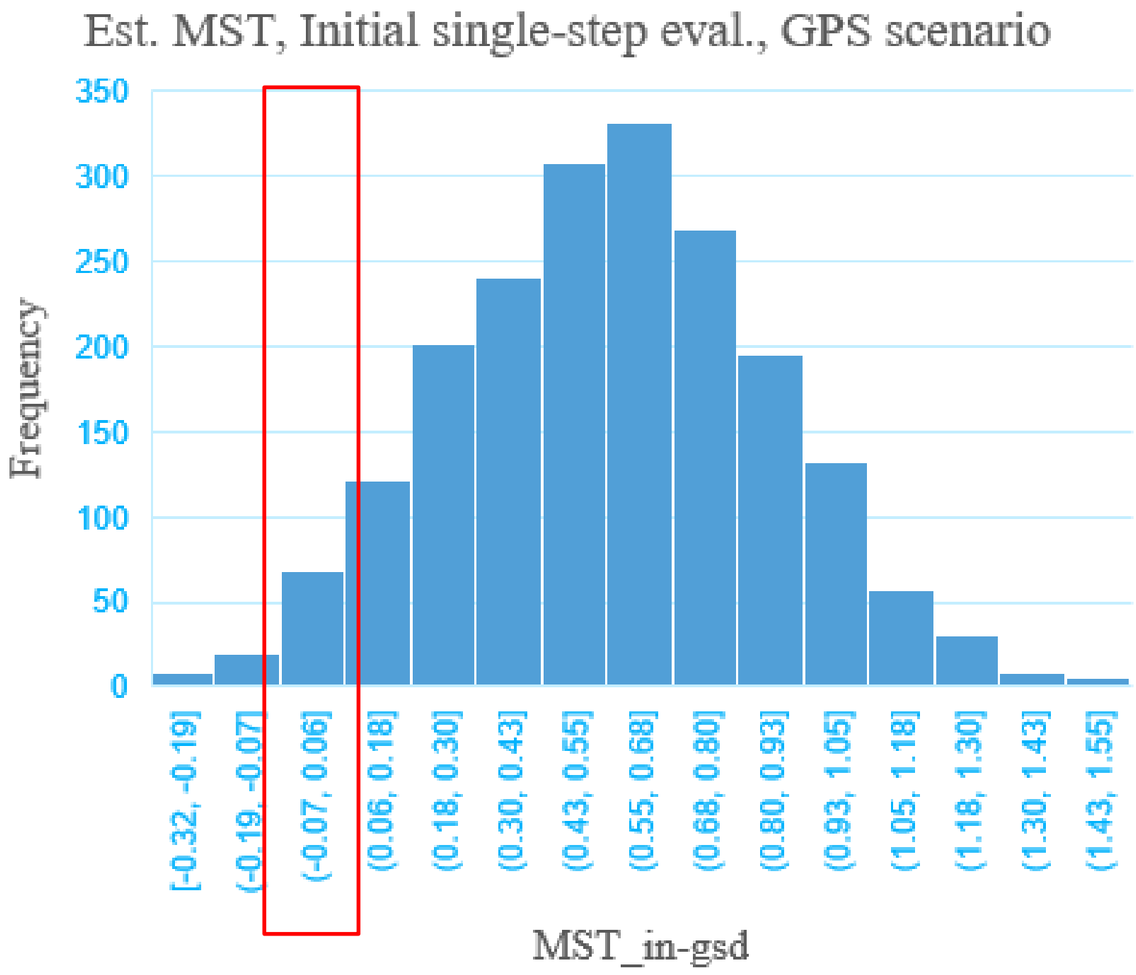


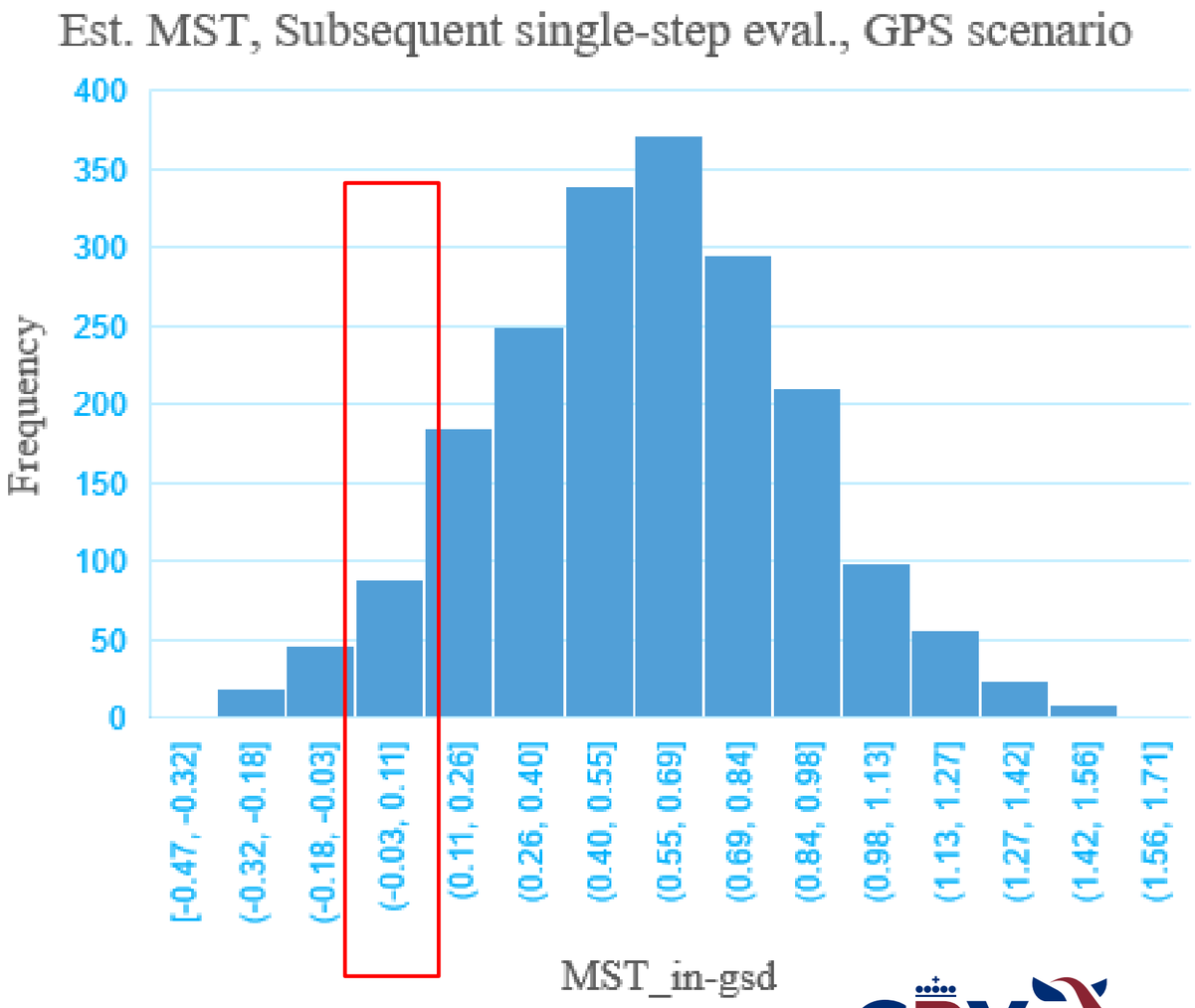


WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Distributions of Mendelian sampling tems

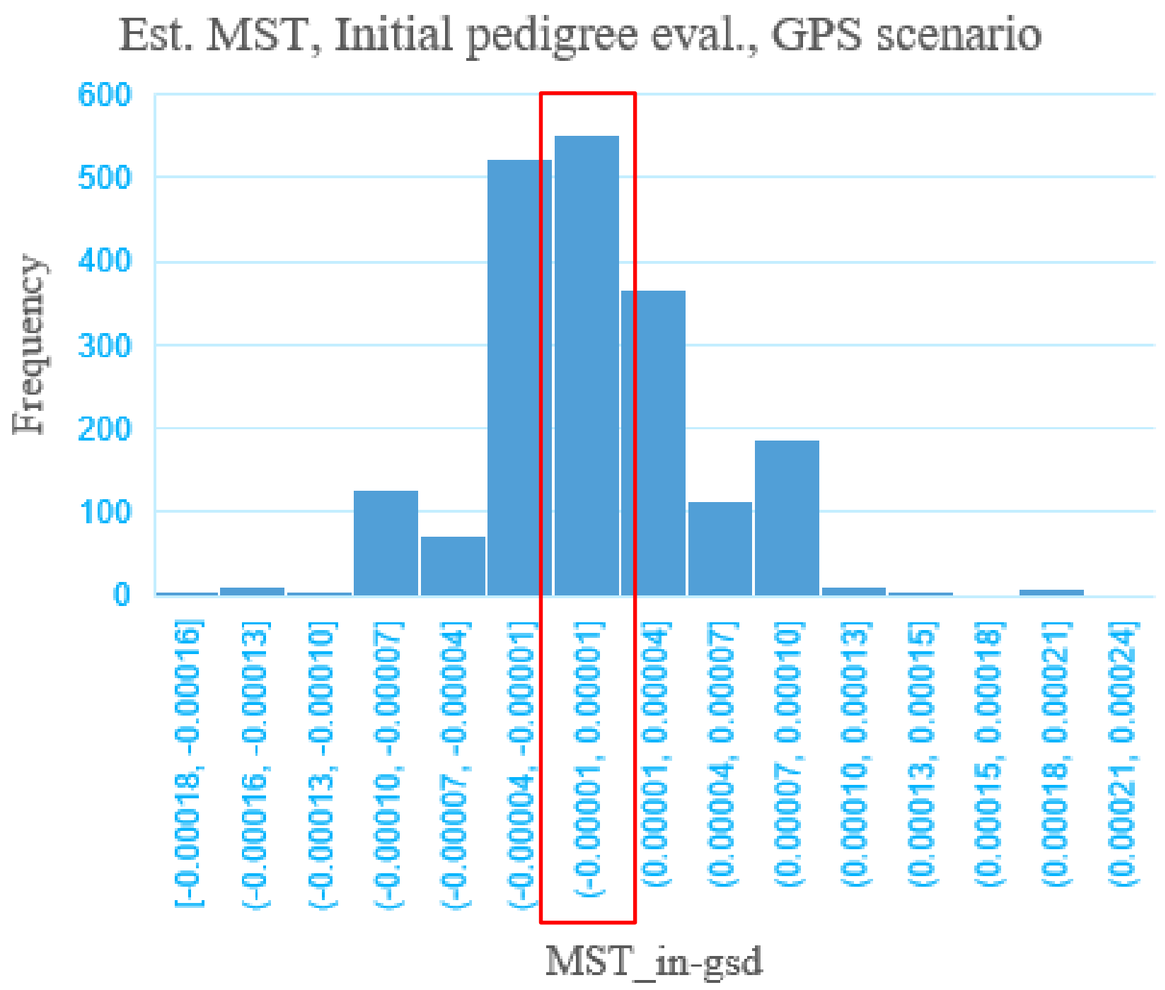


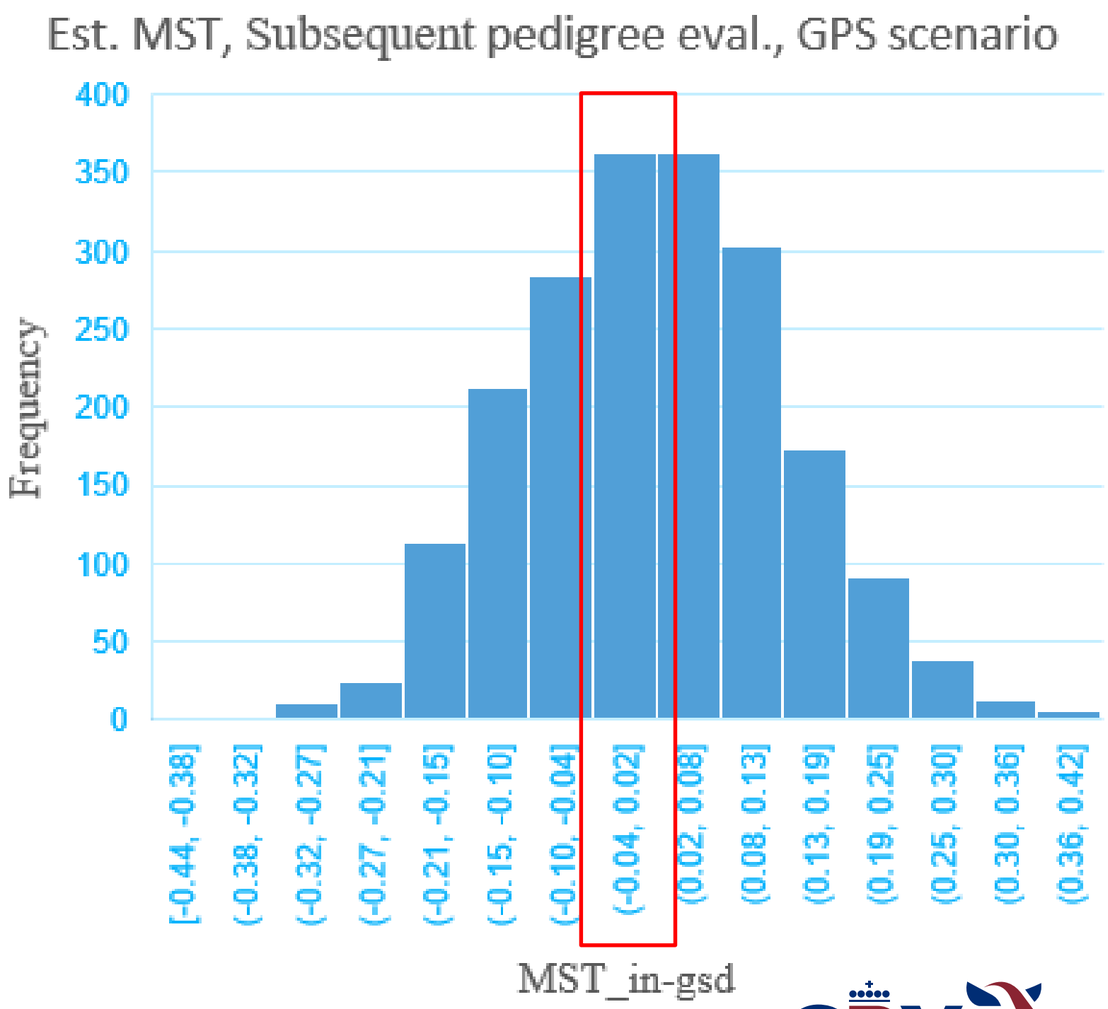


WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Distributions of Mendelian sampling tems





WAGENINGEN UNIVERSITY & RESEARCH

CRV Coöperatie BETTER COWS > BETTER LIFE

Take home message

Preselection bias is not an issue in single-step evaluations!

Acknowledgements

My other PhD supervisors

- prof. dr. ir. Roel Veerkamp
- dr. ir. Jan ten Napel
- dr. ir. Jeremie Vandenplas

My PhD sponsors

BREED4FOOD

WAGENINGEN UNIVERSITY & RESEARCH

Cobb®

CRV
BETTER COWS > BETTER LIFE

HENDRIX GENETICS

Topigs Norsvin
PROGRESS IN PIGS

WAGENINGEN UNIVERSITY & RESEARCH

Genomic preselection in single-step evaluation

Ibrahim Jibrila, Mario Calus, Gerben de Jong

Interbull Technical Workshop, 15/03/2023, Rome

WAGENINGEN UNIVERSITY & RESEARCH

INTERBULL

CRV Coöperatie

BETTER COWS > BETTER LIFE

Using genetic regressions for genomic preselection effects

Pete Sullivan (Lactanet, Canada)
Esa Mäntysaari (Luke, Finland)
Gerben deJong (CRV u.a., Netherlands)

GPS-AI bulls in MACE

- MACE uses biased input EBV, generated without genotypes and therefore ignoring Genomic Pre-Selection (GPS) effects on the Mendelian sampling distributions of most recent AI bulls

Objectives:

1. Develop a **GPS-MACE** international model that accounts for these GPS effects without requiring any genotypes, intending to
2. Reduce bias in future MACE proofs that can still be used as phenotypic input data for national genomic evaluation systems

Genetic regressions for GPS

- We wish to estimate selection effects on **GPS groups** of AI bulls.
- Pre-selection groups (CouSel) based on Country of registration
  - 840+USA are combined, DNK+FIN+SWE are combined as DFS
- To **avoid small groups**, we fit regressions on YEAR by CouSel
  - Estimating trends in GPS (YEAR as a covariable) for each COUNTRY
- To *allow non-linearity*, to *reduce fluctuating* estimates over time, and for *stable estimates* on most recent bulls, we use 3-year knotted linear slopes (in parameter vector **s**) with the following assumptions:

Genetic regressions for GPS

- Assumptions about GPS of dairy sires:
  1. GPS *level = x* and *trend=0* in most recent time period (2014-2017)
  2. GPS *level = 0* and *trend=0* prior to the start of GPS (1980-2008)
  3. GPS *trends during intermediate periods* (2009-2011 and 2012-2014) capture evolving GPS intensities, as the levels go *from 0 to x*
  4. *x = 0 for smallest populations* where *x* cannot be estimated reliably
- Allows for different timings of GPS implementation, and different yearly intensities of pre-selection, for each trait-country combination
- National input data drive all GPS estimations and EBV adjustments

GPS-MACE model

- Current MACE: $y = \mu + Q_1 g + \mathbf{a} + e$
- Current MACE: $y = \mu + (Q_1 g + \mathbf{PA}) + (\mathbf{MS}) + e$
- **GPS-MACE:** $y = \mu + (Q_1 g + \mathbf{PA}) + (\mathbf{Q_2 s} + \mathbf{m}) + e$

GPS effects

GPS effects

$$\overline{\mathbf{MS}} = \mathbf{Q_2 s}, \qquad \bar{\mathbf{m}} = \mathbf{0}$$

GPS-MACE equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{DX} & \mathbf{X}'\mathbf{DZ} & \mathbf{X}'\mathbf{DZQ_2} \\ \mathbf{Z}'\mathbf{DX} & \mathbf{Z}'\mathbf{DZ} + \mathbf{W} \otimes \mathbf{G}_t^{-1} & \mathbf{Z}'\mathbf{DZQ_2} \\ \mathbf{Q_2}'\mathbf{Z}'\mathbf{DX} & \mathbf{Q_2}'\mathbf{ZDZ} & \mathbf{Q_2}'\mathbf{Z}'\mathbf{DZQ_2} + c\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{Q_1g} + \mathbf{a} \\ \mathbf{s} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Dy} \\ \mathbf{Z}'\mathbf{Dy} \\ \mathbf{Q_2}'\mathbf{Z}'\mathbf{Dy} \end{bmatrix}$$

Covariables in **s** have Incidence Matrix: $\mathbf{ZQ_2}$

We can add a Ridge-regression factor: c

$$\mathbf{EBV} = \hat{\boldsymbol{\mu}} + \mathbf{Q_1}\hat{\mathbf{g}} + \hat{\mathbf{a}} + \mathbf{Q_2}\hat{\mathbf{s}}$$

MiX99 Instructions for GPS-MACE

MACE (Example 3-country model)	GPS-MACE (fixed regressions)	GPS-MACE (ridge regression: c=100)
INTEGER **An Cou** REAL **Y D # Y=drp D=edc*R-inv** MISSING **-9999.0** DATAFILE **st-am.data** PEDIGREE **G am+p 1** PARFILE **st-am.para # V(reg)=G, R=1** TABLEFILE **identity_matrix** TABLEINDEX **Cou** MODEL **Y = Cou G(t1 t2 t3| An) ! WEIGHT=D**	INTEGER **An Cou** REAL **Y D # Y=drp D=edc*R-inv** MISSING **-9999.0** DATAFILE **st-am.data** PEDIGREE **G am+p 1** PARFILE **st-am.para # V(reg)=G, R=1** TABLEFILE **identity_matrix** TABLEINDEX **Cou** **# 2 regressions per country = 6 total** REGMATRIX **FIXED yc FIRST=2 LAST=7** REGFILE **ZQ2_incidence** MODEL **Y = Cou G(t1 t2 t3| An) ! WEIGHT=D**	INTEGER **An Cou** REAL **Y D # Y=drp D=edc*R-inv** MISSING **-9999.0** DATAFILE **st-am.data** PEDIGREE **G am+p 1** PARFILE **st-am.para # V(reg)=G, R=1** TABLEFILE **identity_matrix** TABLEINDEX **Cou** **# 2 regressions per country = 6 total** REGMATRIX **RANDOM yc FIRST=2 LAST=7** REGFILE **ZQ2_incidence** REGPARFILE **s_ridge_100** MODEL **Y = Cou G(t1 t2 t3| An) ! WEIGHT=D**

GPS effects accumulate over time

- **Q:** Is Pre-Selection of AI bulls on MS=(GEBV-PA) or on GEBV ?
  1. PA=Between Family: only bulls from the best families are used in AI
  2. MS=Within Family: only the best young bulls within a selected family
- PA (family) pre-selection is based on 2 sources of information
  1. Contribution from traditional EBV of parents
  2. s_{PA} from additional Genomic Information ***on ancestors*** (GEBV-EBV)
- MS (within) pre-selection is based on only the 2nd source of info
  2. s_{MS} from additional Genomic Information ***for the young bull*** (GEBV-PA)

Q_2 includes GPS of ancestors

- The true Breeding Value of a genomic young bull includes his within-family selection ($\mathbf{s_{MS}}$) plus accumulated GPS of his ancestors ($\mathbf{s_{PA}}$)
- Matrix Q_2 links each animal to the sum of these two terms:

$$(\mathbf{Q}_{2:animal} * \mathbf{s}) = \mathbf{s}_{animal} = (\mathbf{s_{PA}} + \mathbf{s_{MS}})$$

$$\mathbf{Q}_{2:animal} = \mathbf{Q}_{2:PA} + \mathbf{Q}_{2:MS}$$

$$\boxed{\mathbf{Q}_{2:animal} = \mathbf{0.5} * (\mathbf{Q}_{2:sire} + \mathbf{Q}_{2:dam})} \boxed{+ \mathbf{Q}_{2:MS}}$$

Q_2 - GPS

$$\boxed{\mathbf{Q}_{1:animal} = \mathbf{0.5} * (\mathbf{Q}_{1:sire} + \mathbf{Q}_{1:dam})}$$

$$\mathbf{Q}_{1:animal} = \mathbf{Q}_{1:PA}$$

Q_1 - UPG
Quaas, 1988

Expected GPS in foreign countries

- We want to estimate GPS effects in the country of selection only:
  - To get only good estimates in an **s** of order = NCOU, rather than estimating [NCOU]*[NCOU] combinations that would include many poor estimates
- We include genetic regressions of GPS effects to foreign scales in matrix Q_2:

$$\mathbf{s}_{animal:\mathrm{B,A}} = \left(\frac{G_{BA}}{G_{AA}}\right) * \mathbf{s}_{animal:A,A}$$

$$\mathbf{Q}_{2:ms:\mathrm{B,A}} = \left(\frac{G_{BA}}{G_{AA}}\right) * \mathbf{Q}_{2:ms:A,A}$$

Step 2 | **Step 1**

Testing the GPS-MODEL

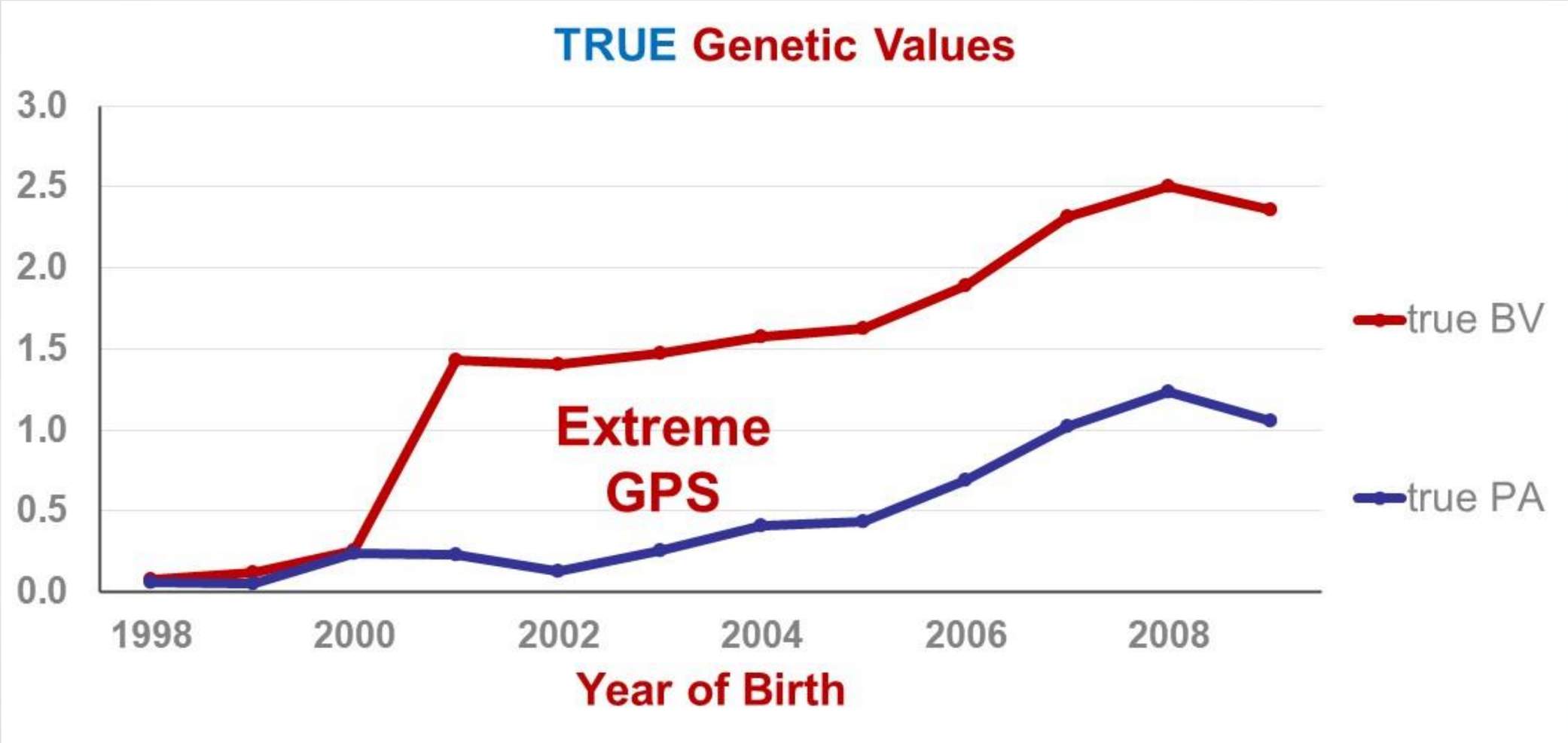
1. **Simulation study**: *unbiased national EBV* input for MACE
   1. A simple design with GPS practiced in only one country
   2. Expectation of MACE output that is unbiased, which is easily tested
2. **Official data study**: *biased national EBV* input used in MACE, after years of GPS in many countries, but with GPS effects not properly included in the national EBV computed without genotypes

1. Simulated Data

- Simulated phenotypes based on observed distributions of PA and MS for proven bulls in the April 2014 MACE evaluation for Protein
  - Youngest proven bulls were born in 2008/2009 (before GPS started)
- GPS effects were simulated as an increase of approximately 1 genetic SD in true genetic means, for the GPS bulls born between 2001-2009, registered and with a national EBV from MACE country #1
  - GPS means were added to the de-regressed EBV used in MACE
  - Input data for MACE were “unbiased” (GPS effects included in DRP)
    - Expectation that GPS effects are properly estimated with a correct model

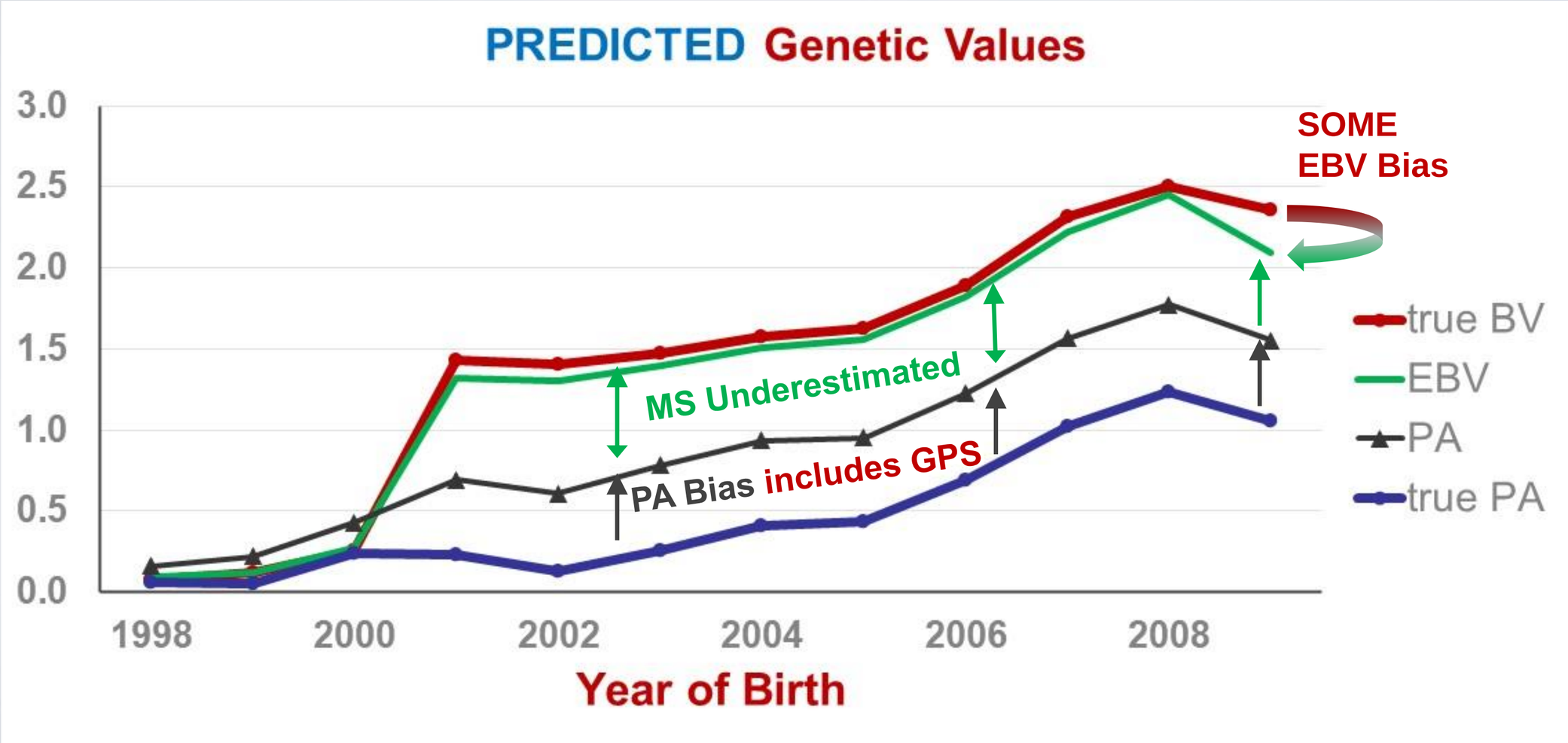
Simulated Data with strong GPS

(Tyrisevä, 2018$_{JDS}$; Benhajali, 2019$_{IB}$)



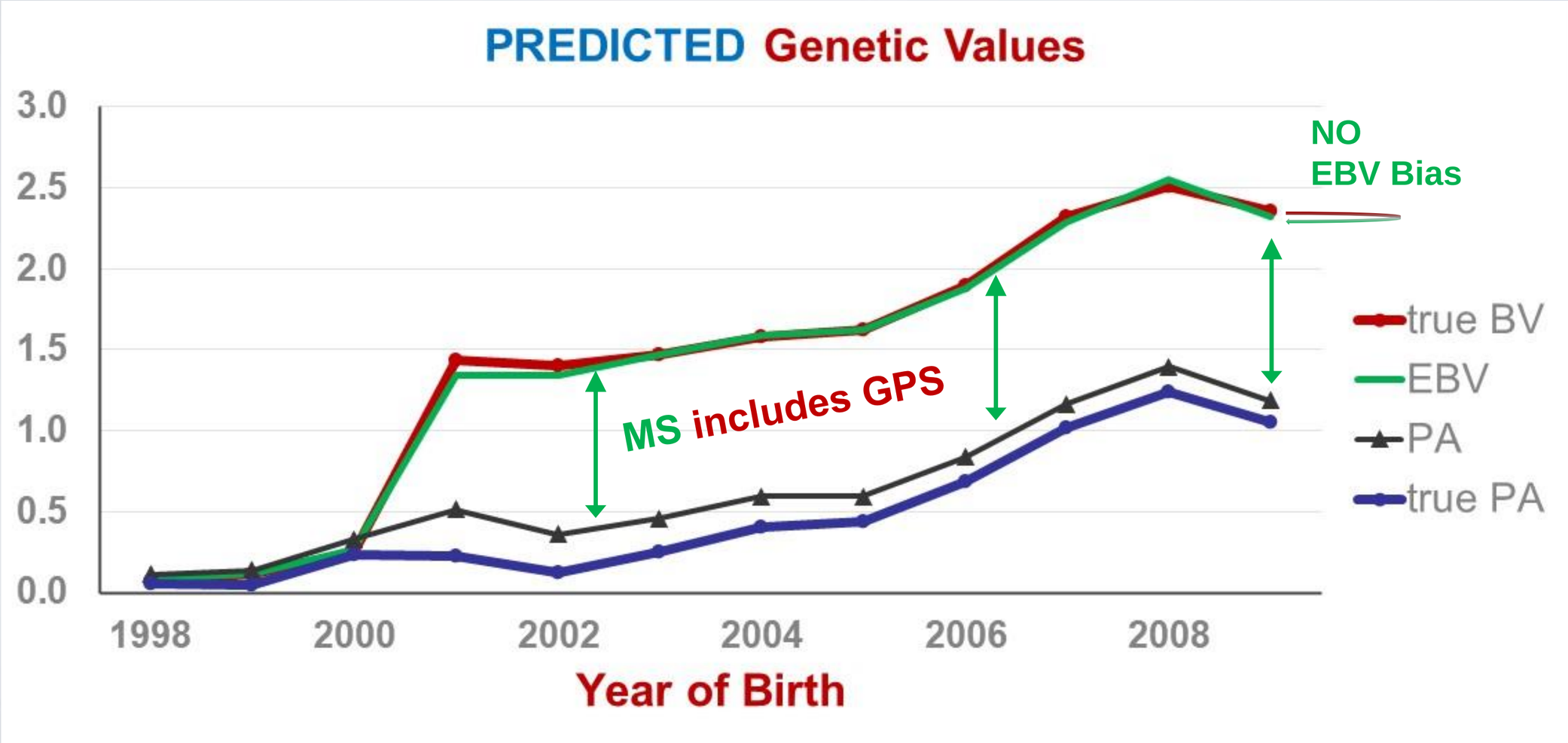
Simulated Data with strong GPS

(MACE with unbiased EBV input)



Simulated Data with strong GPS

(GPS-MACE with unbiased EBV input)



2. Official MACE Data

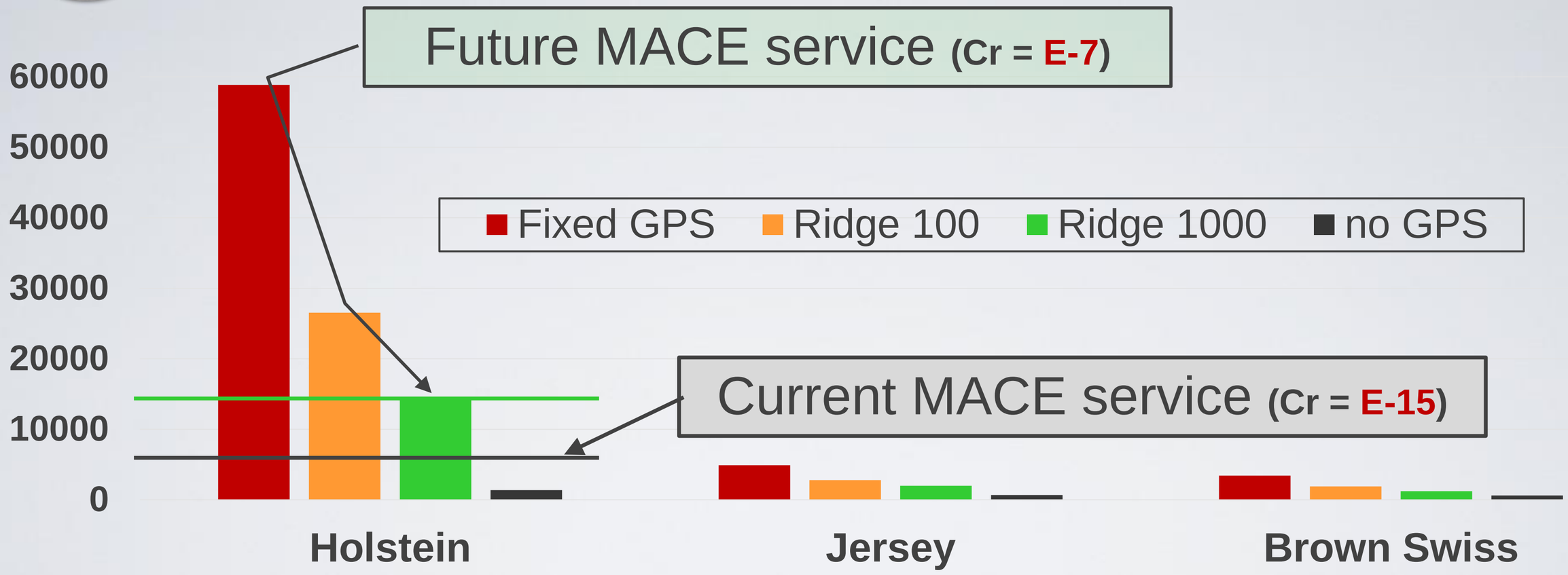
- Official input data used for MACE in April 2022 for:
  - Eight traits: pro, fat, ocs, ous, scs, cc1, int, msp
  - Three “genomic” breeds: Holstein, Jersey, Brown Swiss
- Proven bulls were born as recently as 2017, with approximately 8 completed years of progeny-proven GPS bulls (2009-2016)
- National EBV are biased (i.e. with estimated MS effects that are too small) due to the requirement of ignoring genotypes
  - Expecting GPS effects to be “underestimated” from these data

Results and Discussion

1. Some practical considerations for solving the model
2. Estimates of GPS effects (all on standardized bull proof scales)
3. Impacts of adding GPS effects on the EBV and PA
4. Plans and timeline for implementing GPS-MACE

★ Ridge Regression is FASTER

PCG Iterations to converge (Cr = E-7)



MACE for Protein, April 2022 data

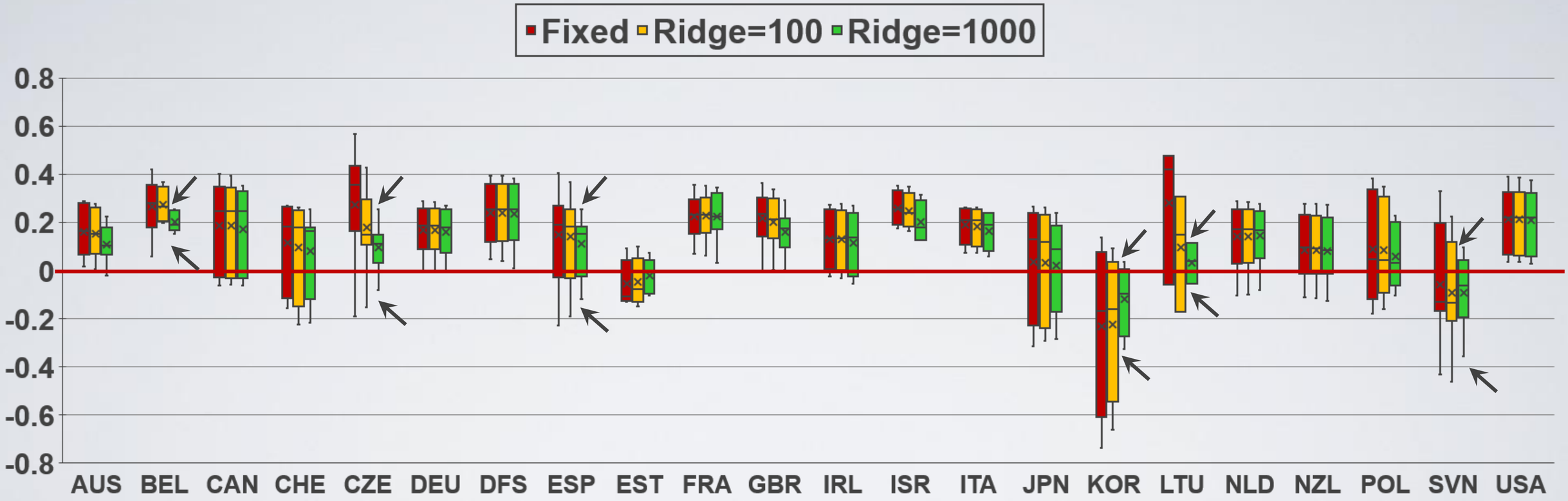
Benefits of Ridge Regression

- GPS-MACE is a more complicated model
  1. We are adding another partition for ANIMAL with $\mathbf{s}_{MS}$
  2. We now estimate selection effects at both ends of the pedigree
     - UPG in the base population and $\mathbf{s}_{MS}$ in the current population
- We are increasing co-linearities and confounding among estimates, and the potential for linear dependencies (i.e. singular equations with no unique solutions) if we treat covariables in **s** as fixed effects
- Fitting Ridge/Random **s** breaks any mathematical dependencies, guaranteeing unique EBV solutions, shrinking V(estimates) and reducing the likelihood of over-fitting the data, to improve “future (i.e genetic) prediction”

★Ridge Estimates are BETTER

Current levels of GPS across 8 Traits

(s_{MS} estimates for Holstein)

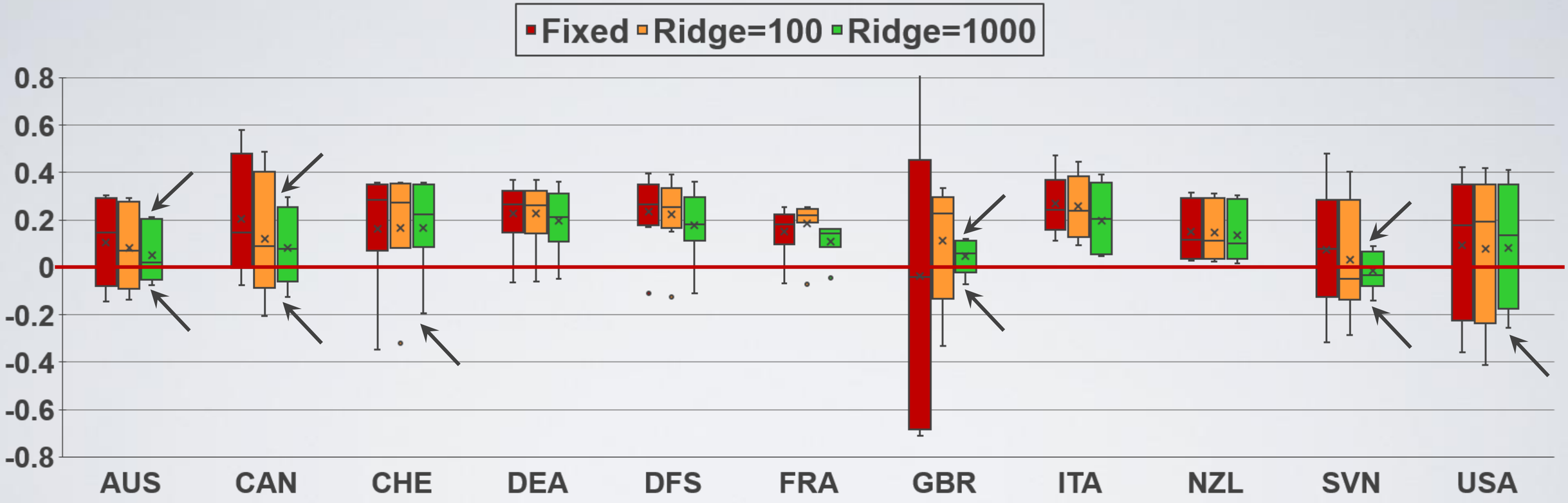


HOLSTEIN - Registered AI bulls born 2014-2017

★Ridge Estimates are BETTER

Current levels of GPS across 8 Traits

(s_{MS} estimates for Jersey and Brown Swiss)

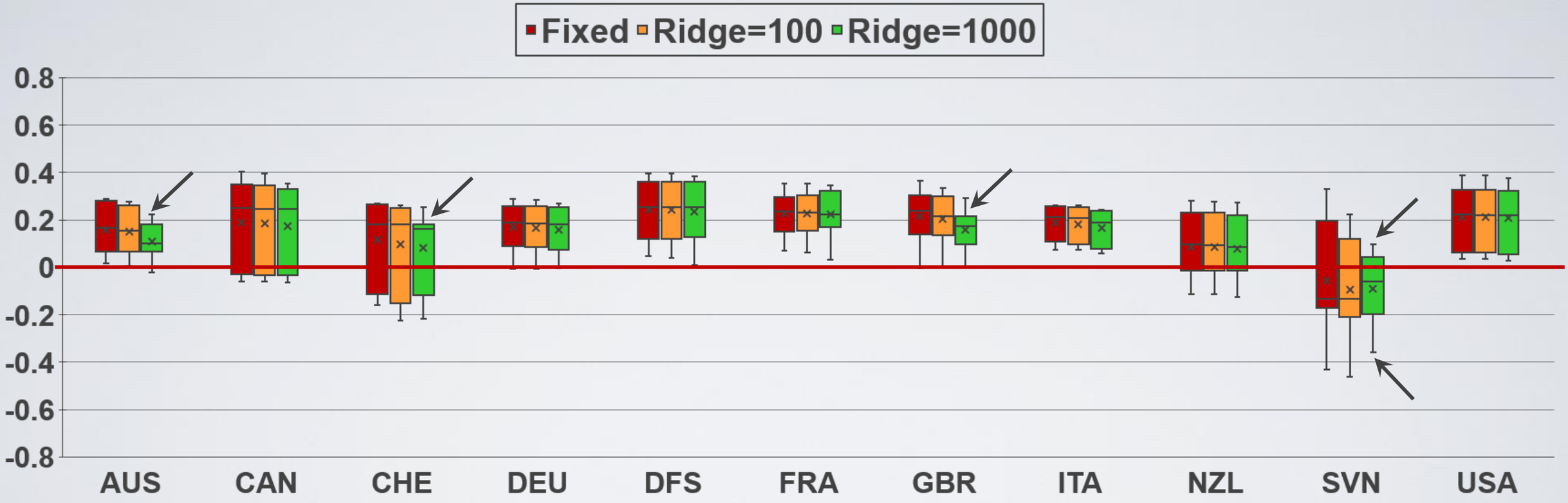


JERSEY and BROWN SWISS - Registered AI bulls born 2014-2017

★Ridge Estimates are BETTER

Current levels of GPS across 8 Traits

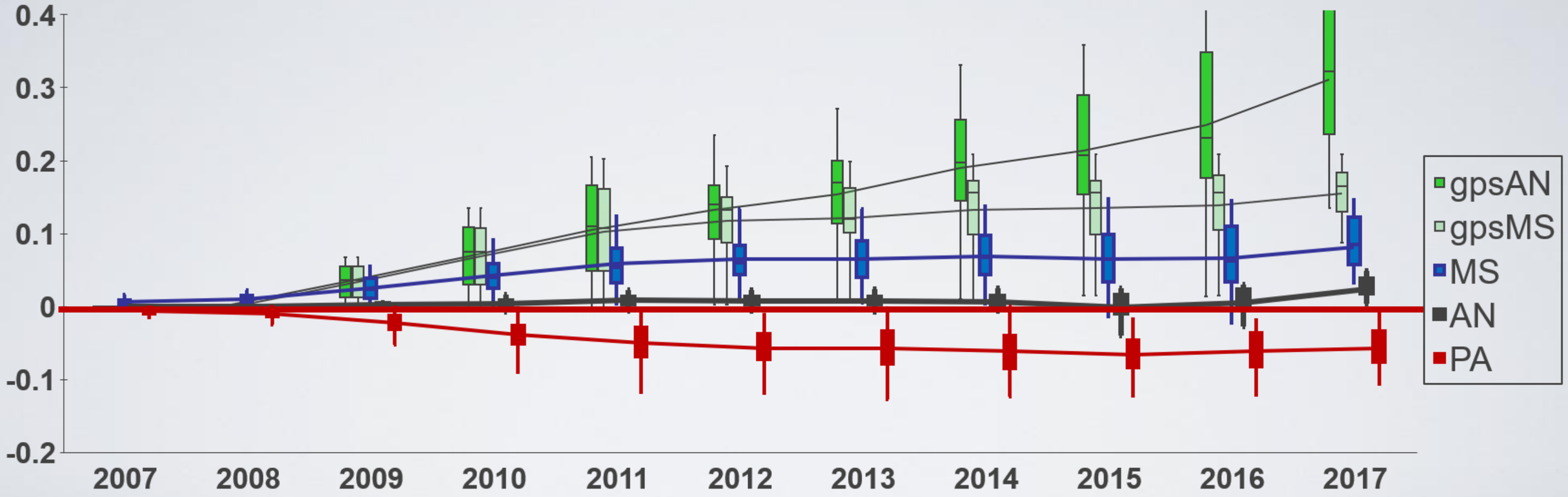
(s_{MS} estimates for Holstein)



HOLSTEIN - Registered AI bulls born 2014-2017

Impact of GPS on MACE proofs

Distributions of Averages by Country of Registration, for the Holstein trait Protein in Canada



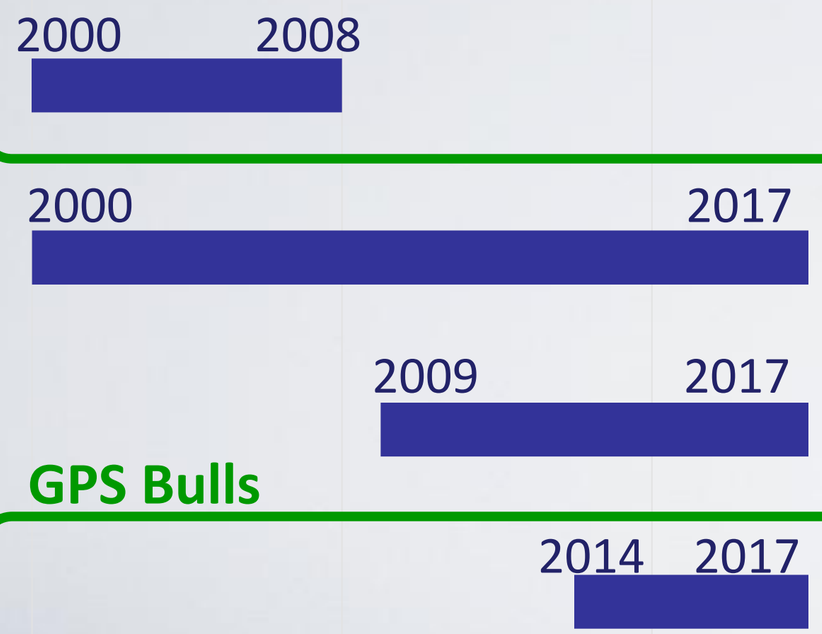
Birth Year of HOLSTEIN bulls EBV-proven in any country

Bull proofs from GPS-MACE vs. MACE

Across all Scales of Evaluation for 8 traits

Scales*Traits: (**176** for Holstein) (**143** for Jersey + Brown Swiss)

AI Sire Birth Year Range



Proof Regressions	Holstein		Jersey and Brown Swiss	
(y=GPS-MACE)	Minimum	Maximum	Minimum	Maximum
Correlation	1.000		1.000	
Slope	0.997	1.004	0.998	1.013
Correlation	0.999		0.998	
Slope	0.993	1.010	1.000	1.032
Correlation	0.997		0.995	
Slope	0.996	1.011	0.995	1.037
Correlation	0.994		0.990	
Slope	0.996	1.021	0.986	1.050

Implementing GPS-MACE

- Expecting ***small EBV changes initially***, for MACE of proven bulls
  - The future MACE proof ***changes will be bigger*** with ***improved national input*** data
- Can immediately expect ***larger changes in PA*** from MACE, which are used directly in GMACE for the young genomic bulls
  - Impacts on ***GMACE results have not been examined yet***
  - The national GEBV - MACE_PA will be larger with GPA_MACE, so the national GEBV should have relatively larger impacts on GMACE proofs for the young bulls
- Implementation of GPS-MACE in Interbull systems could be ready soon
  - A GPS-MACE ***pilot run could be possible*** as early as this fall, 2023

Summary

- GPS effects alter the distributions of GEBV, with effects on both the PA (between-family) and MS (within-family) portions of an AI sire's GEBV
  - BLUP handles most of the PA selection effects, but none of the MS pre-selection
  - GPS effects can be added as an additional term in the model to estimate:
    - Genomic pre-selection on MS of young genotyped bulls
    - Plus any additional PA selection beyond PBLUP, based on additional (GEBV-EBV) information from genotypes of ancestors, which was not picked up already as parental BLUP selection
    - Regressions of GPS effects on time, by country of selection works well (simulated + real data)
  - GPS-MACE programs have been developed for use by Interbull
    - Solve-time is longer than regular MACE, with a more complicated model, but still feasible
    - Programs have been tested on Interbull data and computing systems

Interbull Centre Staff Acknowledgements

- Thanks for organizing working group meetings and communications
- Haifa Benhajali (2017-2018)
  - Initial R&D and programming for GPS simulation and modeling
- Simone Savoia and Marcus Pederson (2019-2021)
  - Transfer and access to Haifa's data and programs, ITBC computing resources, etc.
- Valentina Palucci (2021-ongoing)
  - Collaboration towards a routine implementation of GPS-MACE
  - New processes to incorporate GPS-MACE proofs into GMACE

Questions to the Audience

1. Is additional R&D required before a PILOT run?
   - Adjustments for GPS effects on *V(m)* (HV-GPS-MACE) … do this first?
   - Impacts on *GMACE* results (e.g. with new PA input) … check this first?
   - Should PILOT be ASAP for involvement of national GE centres ?
2. How to *CREATE better* national input data for GPS-MACE?
   - Reducing bias in MACE input data, by properly including “GPS group effects” but not the “individual genotype effects”, has large expected benefits
   - Implementation of GPS-MACE means Interbull would be ready to receive better input
3. How to *VALIDATE if it really is better* national input data for GPS-MACE?

Pre-selection approaches
or
some models with Mendelian sampling terms

Ismo Strandén & Esa Mäntysaari

2/2023

Some background

- The input phenotypes for MACE are derived from EBVs:
  these are biased due to not including genomic based selection decisions.

→ EBVs ignore genomic pre-selection (GPS)

→ EBVs deviate from the expected the more generations genomic selection has been applied.

→ GPS affects MS terms: stronger is selection, larger is E[MS], smaller is Var[MS]

Can a model with Mendelian sampling terms instead of EBV be used to

- compute equivalent breeding values
- lessen the bias in predictions by pre-adjustment of the Mendelian sampling variance

This presentation

- Presents 2 models with the Mendelian sampling terms as unknowns
- Test that these models work on a small MACE data
- Present a possible approach for Mendelian sampling adjustment

Models with Mendelian sampling (MS) term

Standard **BLUP**: $\mathbf{y} = \mathbf{1}\,\mu + \mathbf{Z}\,\mathbf{u} + \mathbf{e}$, where $\mathbf{u} \sim N(\mathbf{0},\mathbf{A}\sigma_u^2)$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$

Expressing **A** by its LDL decomposition: $\mathbf{A}=\mathbf{L}\,\mathbf{D}\,\mathbf{L}'$ allows two equivalent models with an MS term

- **MS I**: $\mathbf{y} = \mathbf{1}\,\mu + \mathbf{L_o}\,\mathbf{m_o} + \mathbf{e}$, where $\mathbf{m_o} \sim N(\mathbf{0},\mathbf{D_o}\sigma_u^2)$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$
  where the subscript o refers to the **A** matrix of the individuals with observation and $\mathbf{A_o}=\mathbf{L_o}\,\mathbf{D_o}\,\mathbf{L_o}'$ (i.e., **LDL** of a subset of **A**).

- **MS II**: $\mathbf{y} = \mathbf{1}\,\mu + \mathbf{ZL}\,\mathbf{m} + \mathbf{e}$, where $\mathbf{m} \sim N(\mathbf{0},\mathbf{D}\sigma_u^2)$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$
  which uses **A** of a full pedigree.

Note: $\mathbf{m}_o$ in MS I has only the size of individuals with observation → MME is smaller than BLUP/SM II

Note: $\mathbf{u} = \mathbf{L}\,\mathbf{m}$, i.e., standard BLUP and MS II models can give all the same estimates.

Multi-trait models with Mendelian sampling (MS) term

Standard AM-BLUP: $\mathbf{y} = \mathbf{X}\,\boldsymbol{\mu} + \mathbf{Z}\,\mathbf{u} + \mathbf{e}$, where $\mathbf{u} \sim N(\mathbf{0},\mathbf{G}\otimes\mathbf{A})$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$

where **G** is the genetic covariance matrix for the traits.

- All vectors and matrices are assumed to be for multiple traits
- $\mathbf{A}=\mathbf{L}\,\mathbf{D}\,\mathbf{L}'$ as before

- **MS I**: $\mathbf{y} = \mathbf{X}\,\boldsymbol{\mu} + \mathbf{L_o}\,\mathbf{m_o} + \mathbf{e}$, where $\mathbf{m_o} \sim N(\mathbf{0}, \mathbf{G}\otimes\mathbf{D_o})$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$
  where the subscript o refers to the individuals with observation

- **MS II**: $\mathbf{y} = \mathbf{X}\,\boldsymbol{\mu} + \mathbf{ZL}\,\mathbf{m} + \mathbf{e}$, where $\mathbf{m} \sim N(\mathbf{0}, \mathbf{G}\otimes\mathbf{D})$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$

What are the L matrices?

MS I: the original model is reparametrized to apply only to phenotyped animals:

- **Z** –matrix in standard BLUP is replaced by $\mathbf{L_o}$ matrix i.e. each observation is modelled using ancestor contributions and an MS term.

RelaX2 instructions:

```
input pedigree                # Pedigree input
  file  amped_selected.ped    # Use this file as pedigree file
  record id sire dam          # This input information

input animals                 # A matrix for these animals
  file  MACE_smaller_123_ids
  record id                   # id column

 output overwrite lower amatrix amatrix_MT.txt # A matrix
```

MS II: the original full **A** matrix is used to make the **L** matrix

- **Z**-matrix has ones for phenotyped individuals
  ZL matrix (new design matrix) includes ancestor contributions from (also non-phenotyped) individuals

```
input pedigree                # Pedigree input
  file amped_selected.ped     # Use this file as pedigree file
  record id sire dam          # This input information

input animals                 # L matrix for these animals
  file MACE_smaller_123_ids
  record id                   # id variable locations

output overwrite lmatrix lmatrix_MT.txt # output L matrix
```

The D –matrices are

Diagonal matrices having the variances of MS terms

MS I: **D** matrix from **LDL** decomposition of $\mathbf{A}_o$

MS II: Simple structure can be computed using pedigree:

base population $d_{ii}=1$

one parent known $d_{ii}=¾$

both parents known $d_{ii}= ½$.

Note: The models need variances $\mathbf{G}\otimes\mathbf{D_o}$ or $\mathbf{G}\otimes\mathbf{D}$ that can be easily computed.

These matrices have blocks of $d_{ii}\mathbf{G}$ where d_{ii} is diagonal from $\mathbf{D_o}$ or **D**.

Pilot test of the concept

- Concept was pilot tested using standard MiX99 package
- Input data generated using RelaX2 (minor change to output the **A** and **L** –matrices)
  - And some help programs to make matrices $\mathbf{L_o}$, **ZL**, $\mathbf{G}\otimes\mathbf{D_o}$ and $\mathbf{G}\otimes\mathbf{D}$.
- An old research data from MACE evaluations were used as an example (Tyrisevä, et al. 2011)

Test data: MACE model, 3 countries/traits

```
DEPENDENT VARIABLES:
  TR TR-NAME        N-OBS       MEAN         SD        MINIMUM      MAXIMUM
  ---------------------------------------------------------------------------
   1  dyd_PROT        7028  -0.67590       16.481       -59.134        391.55
   2  dyd_PROT       16734    1.8439       11.905       -39.712        47.086
   3  dyd_PROT        8900   -5.5464       11.638       -48.024        34.878
```

Standard MACE model in MiX99

Multi-trait data presentation

```
DATAFILE   ../MACE_smaller_123_MT.dat
INTEGER    BULL CTRY1 CTRY2 CTRY3
REAL       dyd1 W1 dyd2 W2 dyd3 W3
MISSING    -8192.0

PARFILE    MACE_smaller.var   # Variance component file

PEDFILE    ../amped_selected.ped # Pedigree file
PEDIGREE   BULL am  # Genetics associated with pedigree

TMPDIR ./tmp

MODEL
   dyd1 = CTRY1  -     -    BULL ! weight= W1
   dyd2 =   -  CTRY2   -    BULL ! weight= W2
   dyd3 =   -    -   CTRY3 BULL ! weight= W3
```

Trait group data presentation

```
DATAFILE   MACE_smaller.dat
INTEGER    BULL COUNTRY
REAL       dyd_PROT WEIGHT
MISSING    -8192.0

TRAITGROUP COUNTRY

PEDFILE    amped_selected.ped # Pedigree file
PEDIGREE   BULL am  # Genetics associated with pedigree

PARFILE    MACE_smaller.var   # Variance component file

TMPDIR ./tmp

MODEL
   dyd_PROT(1) = COUNTRY BULL ! weight= WEIGHT
   dyd_PROT(2) = COUNTRY BULL ! weight= WEIGHT
   dyd_PROT(3) = COUNTRY BULL ! weight= WEIGHT
```

A = L D L' matrix summaries

Regression matrices for MiX99

```
REGMATRIX   heterogeneous reg FIRST=1 LAST=31578
REGFILE     MS_I_L0.txt
REGPARFILE  MS_I_D0_3tr.txt
```

L_0 matrix

$G \otimes D_0$ matrix

MS I: L_o for the solver

31,578 rows and columns → ~4GB in real 4.

492,425,057 non-zeros → 49.4% non-zero (so the lower triangle is almost full)

```
REGMATRIX   heterogeneous reg FIRST=2 LAST=66776
REGFILE     MS_II_ZL.reg
REGPARFILE  MS_II_D_3tr.par
```

ZL matrix

G⊗D matrix

MS II: L for the solver

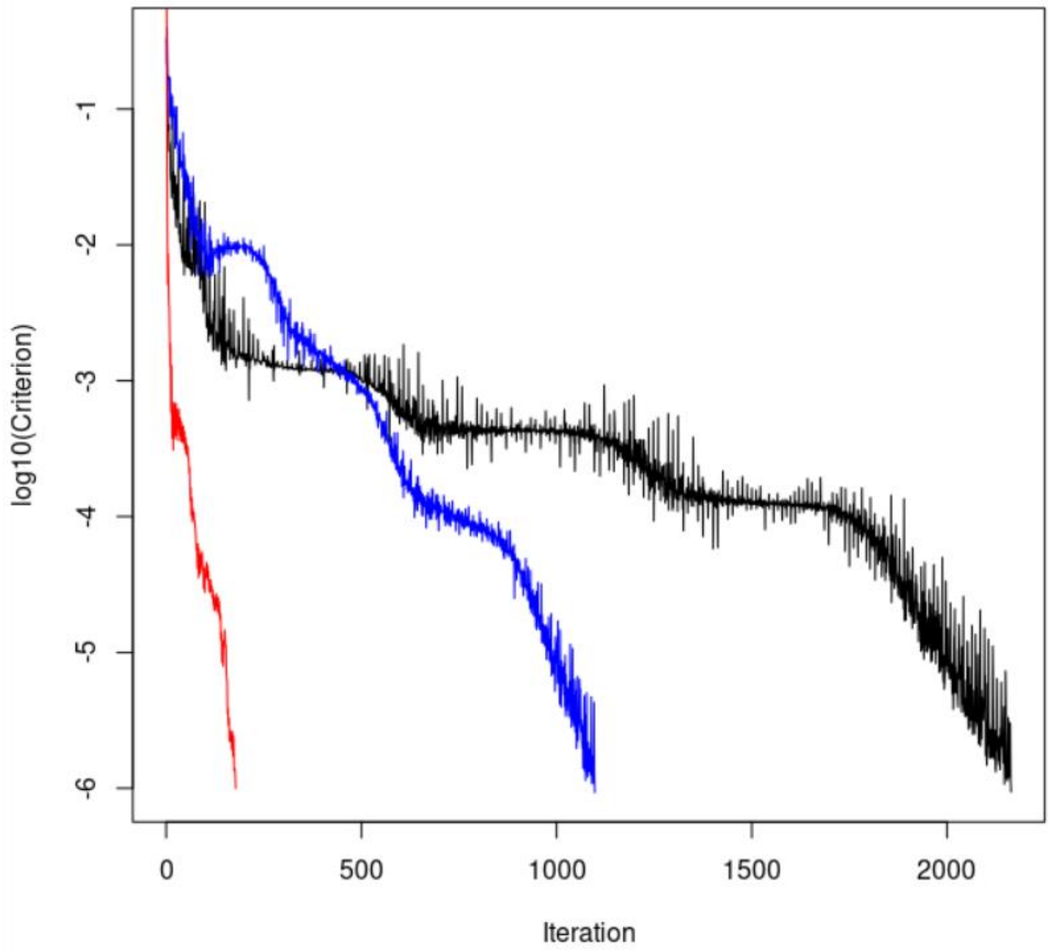
31,578 rows, 66,776 columns → ~8.5GB in real 4.

135776 non-zeros → 0.01% non-zero

MS variances
"diagonal relationship" file

Although both methods can be tested by a regular MME solver, the sparsity pattern in SM II suggests a lower memory use and fast computations can be achieved by using pedigree data and a "half"-Colleau algorithm in the PCG iteration.

PCG Convergence



Red: Standard MACE model, 178 iter.
Black: MS I model, 2,164 iter.
Blue: MS II model, 1,098 iter.

Solver computing times:
Red: 2.5 sec.
Black: 53 min.
Blue: 44 sec.

Convergence criterion: Cr < 10^{-6}

$$C_r = \sqrt{\frac{\left(\mathbf{C}_{\mathrm{MME}}\mathbf{s}^{[k]} - \mathbf{r}_{\mathrm{MME}}\right)'\left(\mathbf{C}_{\mathrm{MME}}\mathbf{s}^{[k]} - \mathbf{r}_{\mathrm{MME}}\right)}{\mathbf{r}_{\mathrm{MME}}'\mathbf{r}_{\mathrm{MME}}}}$$

mix99s -s -p -nocov -cr 1e-6
mix99s -s -p -nocov -srm 1 -cr 1e-6

The MS models showed poor convergence. This is expected (similar to GBLUP vs SNPBLUP).

Further work is needed to improve convergence!

Some thoughts on the possible use of MS model: iterative MS term estimation

- An iterative algorithm can be considered:

1) Solve MS model → solutions for MS term **m**

2) Compute the SD (and average) of **m** within predefined groups

3) Adjust the variance terms in **D** for individuals with deviating **m** using the information in step 2)

4) Go to step 1) with the new **D**, or stop after some rounds.

➔ Highly deviating MS terms are shrunk which may lessen the influence of biased information from relatives.

Summary

- Two models that solve Mendelian sampling terms directly can be used

- Standard software can be used to solve these models, although computationally more efficient algorithms are needed for large data sets.
  - L matrix not given as input but instead solved implicitly from the pedigree

- Convergence of these models was poorer than the standard relationship matrix-based models
  - May have to become a larger issue when more traits (countries) are analyzed

- The Mendelian deviation adjustment algorithm was not fully formulated nor tested. Ideas?

Questions to the Audience

1. Is additional R&D required before a PILOT run?
   - Adjustments for GPS effects on *V(m)* (HV-GPS-MACE) … do this first?
   - Impacts on *GMACE* results (e.g. with new PA input) … check this first?
   - Should PILOT be ASAP for involvement of national GE centres ?
2. How to *CREATE better* national input data for GPS-MACE?
   - Reducing bias in MACE input data, by properly including “GPS group effects” but not the “individual genotype effects”, has large expected benefits
   - Implementation of GPS-MACE means Interbull would be ready to receive better input
3. How to *VALIDATE if it really is better* national input data for GPS-MACE?

Wrap up Session


Interbull Technical Workshop
Thank You