International bull evaluation by GBLUP with prediction population(s)

Ignacy Misztal, Daniela Lourenco,

University of Georgia

Breno Fragomeni, University of Connecticut

Heather Bradford, Virginia Tech

Interbull and genomics

- Before genomics: int'l evaluation based on proven sires and MT sire model

- With genomics
  - marketed mainly young sires
  - In-country evaluations more accurate
  - Now, Interbull source of external data
    - Compromised by genomic selection

- GBLUP MACE possible- if genotypes available
  - APY algorithm for inversion of G

- Idea of int'l MT SNP model (Goddard, 2011)
  - Each country contributes LHS and RHS of SNP BLUP
 - MT model with corr among countries < 1

Problems with SNP model

- Bigger countries not participating
  - Why give away results of large investments

- Attractive for smaller countries but
  - MT SNP model hard to do
  - Each country may have different SNP sets
  - Lots of data (LHS = 50k x 50k)

Genomic information

- SNP effects -what are they?
  - Estimates of causative SNP?
  - Markers to QTL?
  - Tags to important ancestors?

  - Nearly identical prediction with very different SNP chips
  - Indirect estimates of chromosome segments?

- Genomic info in reference population – genotypes + (pseudo)phenotypes
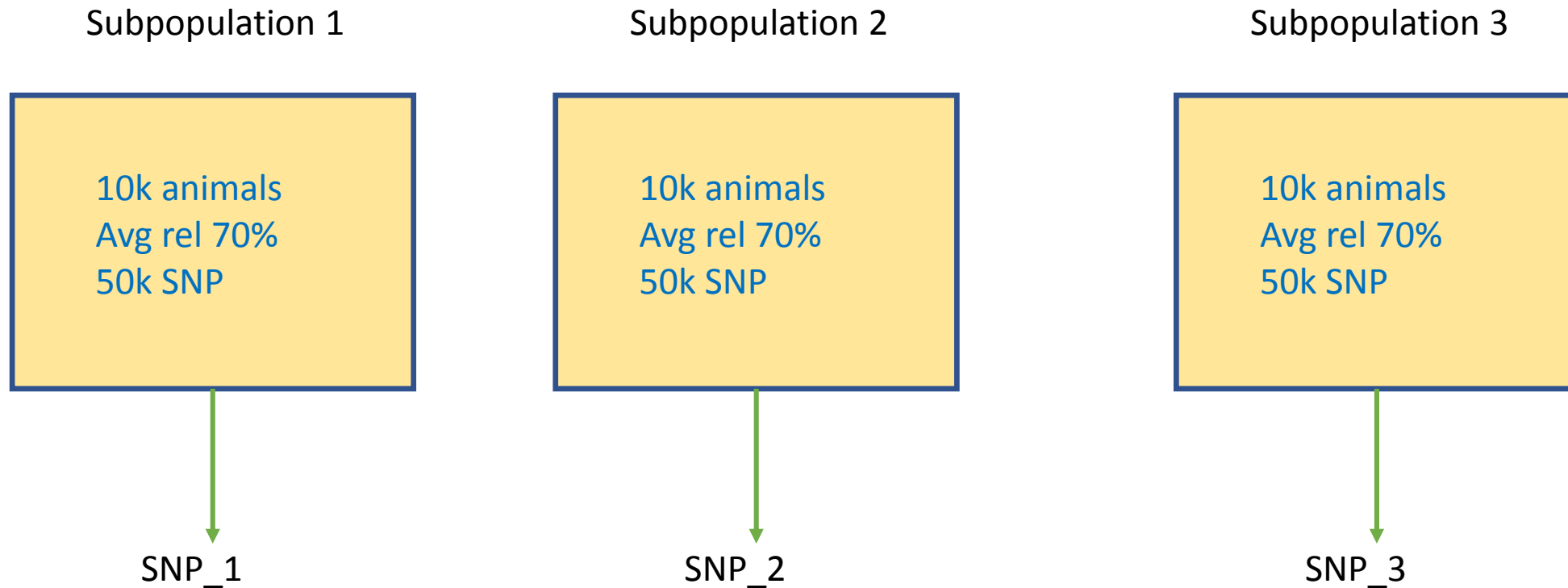
Formulas

$$(D + \alpha G^{-1})\hat{u} = y$$

GBLUP
u = GEBV, y = DRP, α= variance ratio
D = function of reliability

$$\hat{a} = \frac{Z'G^{-1} * \hat{u}}{\sum_{i=1}^{m} 2p_i q_i}$$

Conversion of GEBV to SNP effects
a= SNP effects, Z = genotypes

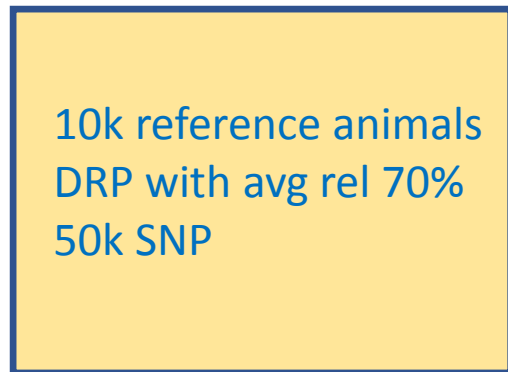$$y^* = (I + \alpha G^{-1})Z\hat{a}$$

Creating pseudo-observations for same SNP solutions
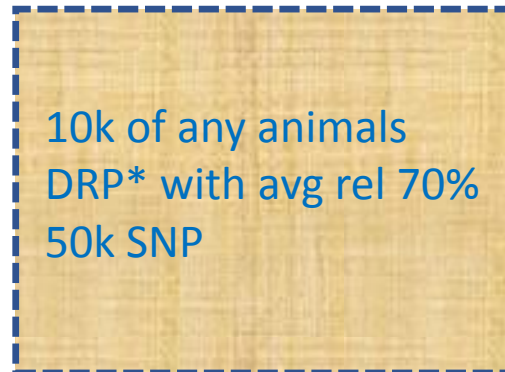y* = DRP for alternative populations

Real and alternative reference population

Reference population

Alternative population

10k reference animals
DRP with avg rel 70%
50k SNP

10k of any animals
DRP* with avg rel 70%
50k SNP

- DRP* - computed for same SNP effects
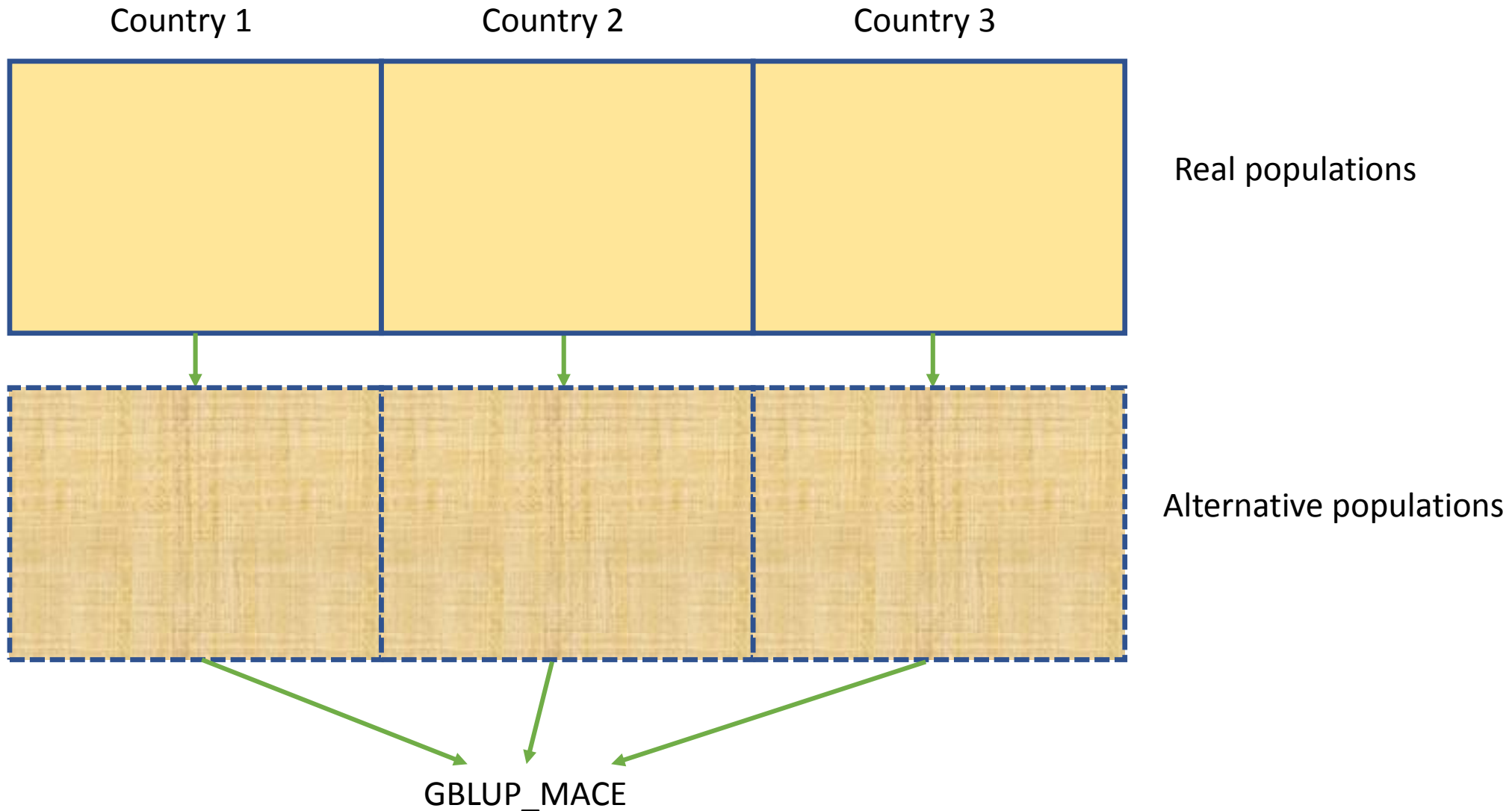
- Any animals – cows, unused bulls, etc.

SNP_ref → SNP_ref

$(I + \alpha G^{-1})GEBV = DRP$ GBLUP

$DRP^* = (I + \alpha G^{-1})Z \widehat{SNP}$ pseudo DRP

$$\widehat{SNP} = \frac{Z'G^{-1} * \widehat{GEBV}}{\sum_{i=1}^{m} 2p_i q_i}$$ Conversion to SNP

GMACE II

Country 1   Country 2   Country 3

Real populations

One trait per country

Alternative population
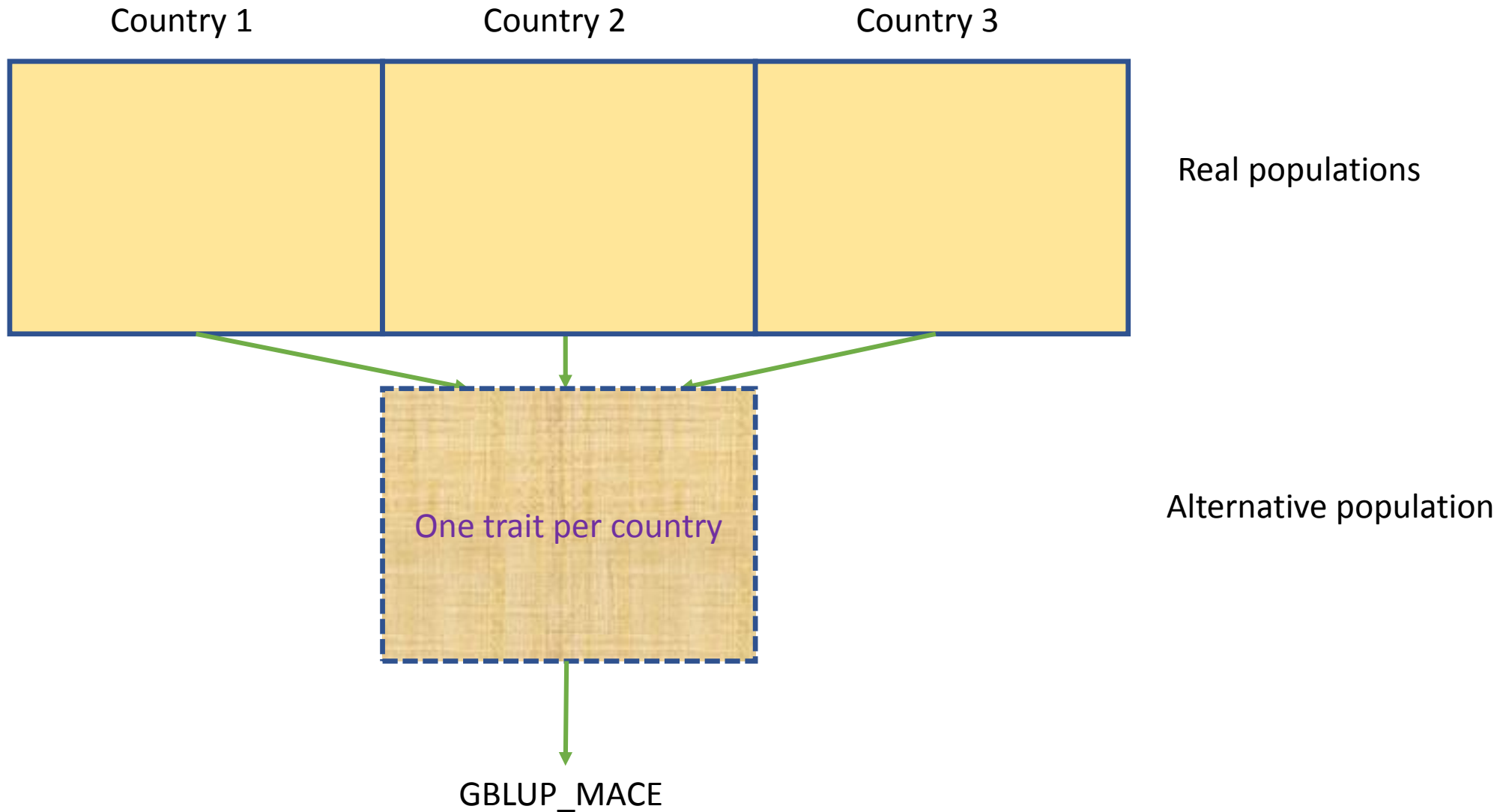
GBLUP_MACE

GBLUP-MACE algorithm

- Compute national SNP effects
- Create (or use Interbull) prediction population
- Create pseudo-phenotypes for prediction population(s)
- Use multiple-trait GBLUP for prediction population(s)
- Convert national GEBV into SNP effects

Simulation outline

30 k animals, 4 traits correlated at 0.8

Traits 1-4

Country 1	Country 2	Country 3	Country 4	Prediction	Validation
Trait 1	Trait 2	Trait 3	Trait 4	DRP1* DRP2* DRP3* DRP4*	Traits 1-4
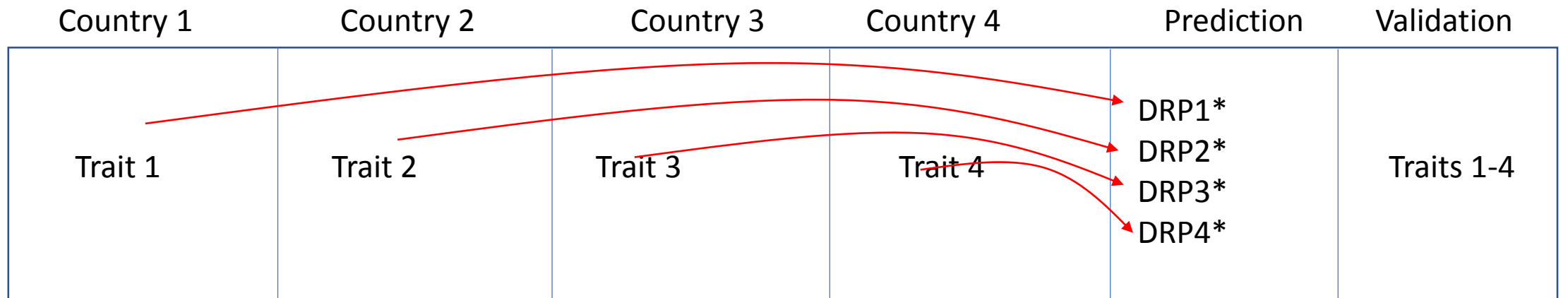
Simulation

- QMSIMM
- $h^2 = 0.25$
- Ne=120
  - 30 males and 20k females/gen
  - 20 generations
  - Last 5 generations with phenotypes
- Mimicking bovine genome
- 500 QTL
- Validation in 5k young animals
  - Artificial population

- 4 traits
  - rel =71%
  - Genetic correlations 0.8

Analyses

- ST-GBLUP – original population within country

- ST-ALTER – alternative population within country

- MT-GBLUP – across country with original populations

- MT-ALTER – across countries with alternative population

Results – 5k animals in each country

	ST-GBLUP	ST-ALTER
COUNTRY1	0.82	0.81
COUNTRY2	0.83	0.83
COUNTRY3	0.84	0.83
COUNTRY4	0.82	0.82

Results – variable number of animals in each country : 5k - 2k - 2k - 1k

	ST-GBLUP	ST-ALTER
COUNTRY1 5K	0.81	0.80
COUNTRY2 2K	0.67	0.66
COUNTRY3 2K	0.68	0.68
COUNTRY4 1K	0.51	0.51

Accuracies with different number of genotyped animals in common

COMMON ANIMALS	ST-GBLUP	MT-GLUP - ST_GBLUP	MT-ALTER - ST_GBLUP
100%	0.84	+0.02	+0.00
50	0.81	+0.04	+0.03
0	0.83	+0.03	+0.02

What prediction population?

1. Interbull prepares a prediction population with size > Me, i.e., > 12,000
   * Genotypes include all SNP used by member countries

2. Each country prepares own prediction population

3. Combination

Comments

- Computational aspects
  - Upgrade MACE by replacing A^{-1} by G_{APY}^{-1}
  - Double counting via by residual correlations

- Why loss of information?
  - Details?
  - Unfinished work?
  - Independent chromosome clusters of highly variable size

- Reluctance to share SNP effects
  - Release only fraction of information
  - Release only for less popular traits
  - Are SNP effects converging?

Advantages of GBLUP MACE

- SNP or GEBV on each country scale

- Everybody gains

- Gains for competition restricted
  - Extra gains possible with local blending and causative SNPs

- Easy implementation

- Future: GBLUP with original data

Conclusions

- GBLUP MACE possible without submitting original genotypes with alternative populations

- Computations simple, some theoretical work needed

- Some loss of information compared to combined prediction with original populations

- Perhaps good compromise

Acknowledgements

- Comments by H. Jorjani, Tom Lawlor, and Paul VanRaden
- Grant from Holsteins Assoc.
- AFRI grant 2015-67015-22936 from USDA NIFA

Shogo Tsuruta

Ignacio Aguilar

Breno Fragomeni

Ivan Pocrnic

Daniela Lourenco

Yutaka Masuda

Andres Legarra

Heather Bradford

Possibilities of SNP model for selected countries

- Submit only SNP effects, # animals and avg reliabilities
- Output: SNP effects on each country scale
- Easy possibility of conversion between different SNP sets
- Simple computations

Assumptions

- Any set of genotyped animals produces asymptotically equivalent SNP effects

N subsets, index i
u_i – breeding values
a_i –SNP effects
Z_i – Gene content

$$(\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{I}\alpha)\,\widehat{\mathbf{a}}_i = \mathbf{Z}_i'\mathbf{y}$$

$$E(\mathbf{Z}_i\,\widehat{\mathbf{a}}_i) = \mathbf{u}_i$$ With large population, estimable functions of SNP effects converge

$$E(\mathbf{Z}_j\,\widehat{\mathbf{a}}_i) = \mathbf{u}_j$$ Estimates from one subset useful to other subsets

Algorithm

- Collect
  - $\widehat{\mathbf{a}}_i$ - SNP effect from i-th country
  - n_i – number of animals used to produce SNP efects
  - rel_i – average reliability

- For each country:
  - Pick up n_i genotypes Z_i from same general population
    - For Holsteins, could be US Holstein cows or bulls
  - Calculate DGV of each animal:

$$\widehat{\mathbf{u}_i} = \mathbf{Z}_i \ \mathbf{a}_i$$

  - Using GBLUP, calculate pseudo-observations y_i
    - β_i function of h² and reliability

$$y_i = (\mathbf{Z}_i \ \mathbf{Z}_i' + \mathbf{I}\beta_i) \ \widehat{u_i}$$

- Run MT GBLUP
- Convert GEBV into SNP effects

Extras

- SNP effects could be for any SNP chip
  - After conversion to DGV, SNP chip is irrelevant

- Conversion back could be to any chip
  - Indirect formulas can use any chip

Questions

- Is algorithm sensitive to choice of genotyped animals?
  - May be more or less related

- Scaling issues for each country

- If low corr of DGV among countries, different traits or models?

Formulas

- For pseudo information

$$y_{ij} = \mu + u_{ij} + e_{ij}, \qquad var(u_{ij}) = \boldsymbol{\sigma_u^2}, var(e_{ij}) = \boldsymbol{\sigma_e^2}$$

$$\boldsymbol{y}_i = (\mathbf{Z}_i \ \mathbf{Z}_i' + \mathbf{I}\beta_i)\, \widehat{\boldsymbol{u_i}} \qquad\qquad \beta_i = \frac{\boldsymbol{\sigma_e^2}}{\boldsymbol{\sigma_u^2}}$$

y is constructed based on assumed variances so any is good provided that it is used later
Can use β=1