Multitrait across country genomic evaluations for EuroGenomics countries

Hanni Kärkkäinen[1], Vincent Ducrocq[2], Zengting Liu[3] & Esa Mäntysaari[1],
[1]Luke, Finland; [2]INRAE, France; [3]vit, Germany

EG SNP MACE Group:
Suzanne De Roo, EuroGenomics; Sander De Roos, CRV; Juan Pena, Conafe

EuroGenomics SNP MACE projects

EuroGenomics: Germany, Nordic countries (Denmark, Finland, Sweden), Netherlands, France, Poland and Spain

- The first EuroGenomics SNP MACE –project 5/2018-5/2020 (EG, Luke and INRAE)
- The goal was to develop multitrait across country SNP BLUP model using shared EuroGenomics bull data directly

After testing the model, it was decided that using the shared bulls is not enough

- Countries want to include full national reference information (cows)
- Without sharing the pheno- and/or genotypes

➡ Meta-analysis using information from national full reference evaluations, not raw data (comp. Savoia, "SNPMace")

EG SNP MACE concept

1. Countries perform genomic evaluations with their own data and method $\Longrightarrow$ National SNP estimates

2. Countries impute the genotypes they use to the common EG SNP density

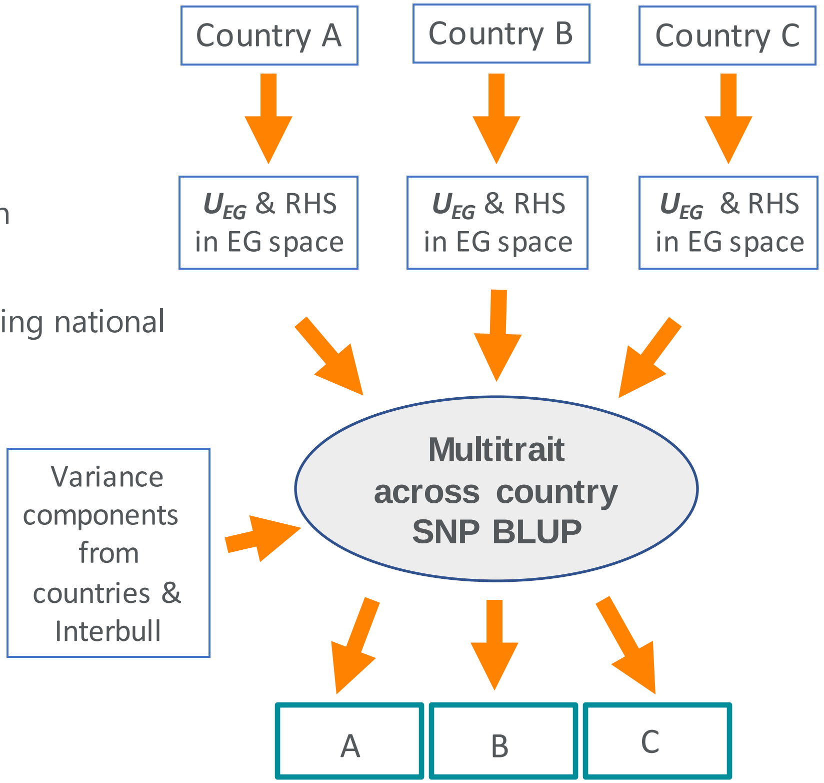
3. EG SNP MACE preprosessor generates blocks of MME using national ***full reference*** and ***national SNP estimates*** compliant with common EG SNP set
$\Longrightarrow$ Pseudo data: RHS_{EG} and $\boldsymbol{U}_{EG}$

4. **Countries share the pseudo data**

5. **Pseudo data is plugged into across country MT SNP BLUP model**
   - Model solved to get SNP-solutions utilizing the full EuroGenomics reference population

6. Results of country specific SNP-solutions are converted back to national SNP set space

National full reference genomic evaluation



Country A

Country B

Country C

$\boldsymbol{U}_{EG}$ & RHS in EG space

$\boldsymbol{U}_{EG}$ & RHS in EG space

$\boldsymbol{U}_{EG}$ & RHS in EG space

Variance components from countries & Interbull

Multitrait across country SNP BLUP

A

B

C

National SNP-effects with EuroGenomics information

Pseudo data

Shared pseudo data comprise ***pivoted Cholesky factorization*** of the national full reference MME LHS, and ***RHS or SNP solution vector***

- Cholesky decomposition = "taking square root" of a matrix
- Contains the same information as the full LHS
- Cholesky matrix can be directly used in standard MME solving programs: cf. $\boldsymbol{U}'\boldsymbol{U} \Leftrightarrow \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}$, where $\boldsymbol{U}$ is the upper triangular of the factorization
- Pivoted Cholesky matrix is always smaller than full LHS
  - size depends on rank of the genotype matrix $\boldsymbol{Z}_{EG}$
  - *e.g.* for a country with 3000 genotyped bulls and no cows, max size of $\boldsymbol{U}_{EG}$ is $3000 \times 50{,}000$

Proof of concept of pseudo data

Pilot tested with shared EG bull data

- Shared data consists of 35,000 observations for protein yield, somatic cell score and female fertility
- 46,342 segregating biallelic SNP genotypes
- Practically the same SNP solutions and DGV as direct usage of raw genotype data
- Correlation between SNP-solutions > 0.99 and between DGV> 0.999
- Within the shared data set, all countries have the same SNP markers

Trait	Country	Correlation between	
		SNP-solutions	DGV
pro	DEU	0.995	1.000
pro	DFS	0.995	1.000
pro	FRA	0.995	1.000
pro	NLD	0.995	1.000
pro	ESP	0.995	1.000
pro	POL	0.996	1.000
scs	DEU	0.996	1.000
scs	DFS	0.996	1.000
scs	FRA	0.996	1.000
scs	NLD	0.996	1.000
scs	ESP	0.996	1.000
scs	POL	0.996	1.000
cc2	DEU	0.998	1.000
cc2	DFS	0.998	1.000
cc2	FRA	0.998	1.000
cc2	NLD	0.998	1.000
cc2	ESP	0.998	1.000
cc2	POL	0.998	1.000

Using full reference population data

Countries use different SNP sets and different models *vs.*
SNP MACE model is based on a single common marker set

1. Establish a common EuroGenomics SNP marker set

2. Full national reference population imputed to the common EG set

3. The LHS of country i can be built directly with the common marker set $\boldsymbol{Z}_{EG_i}$ genotypes

   $$\longrightarrow \quad \mathrm{LHS}_{\mathrm{EGi}} = \boldsymbol{Z}'_{EG_i}\boldsymbol{R}_i^{-1}\boldsymbol{Z}_{EG_i} + \lambda_i\mathbf{I}$$

4. The national marker effect estimates $\hat{g}_i$ are projected on the common marker set to get

   $$\longrightarrow \quad \mathrm{RHS}_{\mathrm{EGi}}$$

Common EG SNP set

- Union of *autosomal*, non-private SNPs the countries use in genomic evaluation, from
  - versions of Illumina 50k chip or
  - public parts of EuroGenomics MD chips

- All DEU, DFS, POL and ESP markers included
  - Some haplotype-related FRA markers excluded

- NLD is currently changing their SNP set
  - Their current markers not considered in building the common set

	DEU	POL	ESP	DFS	FRA	NLD	EG
DEU	44747	44692	44091	43318	41349	9029	**44747**
POL	0.99	45331	44453	43533	41476	9059	**45331**
ESP	0.97	0.97	46161	44878	42446	8980	**46161**
DFS	0.95	0.95	0.97	46341	42897	9089	**46341**
FRA	0.84	0.84	0.85	0.86	53469	8550	**47171**
NLD	0.22	0.22	0.21	0.22	0.19	37995	**9303**
EG	0.94	0.95	0.96	0.96	0.91	0.21	**50112**

Table: On diagonal number of SNP in national (and EG common) set, off diagonal number (above) and proportion (below) of common loci.

The Union EG set includes all DEU, POL, ESP and DFS loci (red).

Imputation to common EG SNP set

Current genotype exchange includes

1. Public part of EG MD chip
2. Illumina v2 and v3

→ These markers are already imputed by countries

Currently the countries

1. Select markers they use in GS
2. Impute selected to full ref population

For EG SNP MACE countries should

1. First impute full EG set markers to full ref population
2. Then select the ones they use in own evaluation

Adding new SNP to a country's current set requires changes in genotype imputing pipeline

→ **Countries need some time to implement the pipeline**

→ Start testing with smaller set = intersection of national sets

Further developments

After the basic model is built and tested, we move into developing the evaluation further:

1. ***Reliability estimation*** for SNP effect solutions / individual animal solutions

2. Inclusion of ***external information*** (non-EG countries) into the evaluations
   - Implementation for this depend on
     i. continuity of current MACE service and
     ii. possible realization of Interbull SNP MACE

3. Include ***residual polygenic effect*** into the model
   - Pedigree based "pseudo markers"
   - Do not require exchange of country estimated individual animal RPG effects

4. Building of the ***evaluation pipeline***

Thank you!

Luke
NATURAL RESOURCES
INSTITUTE FINLAND