Interbull estimation of SNP effects

Michael Goddard

2017

Intro

Sale of semen increasing based on genomic EBVs

We want them to be as accurate as possible

(for all traits in all breeds and countries and between breeds)

i.e. We want estimated SNP effects to be as accurate as possible

High accuracy ← high N, non-linear estimation, one-step, sequence data, functional information

Intro

High N

Human genetics

meta-analysis of 270,000 people for height → more SNPs,
increased accuracy

in UK 500,000 people with WGS

in USA 1M people with WGS

Dairy cattle

1,000,000s world wide if we collaborate

not within-breed, within-country for all traits

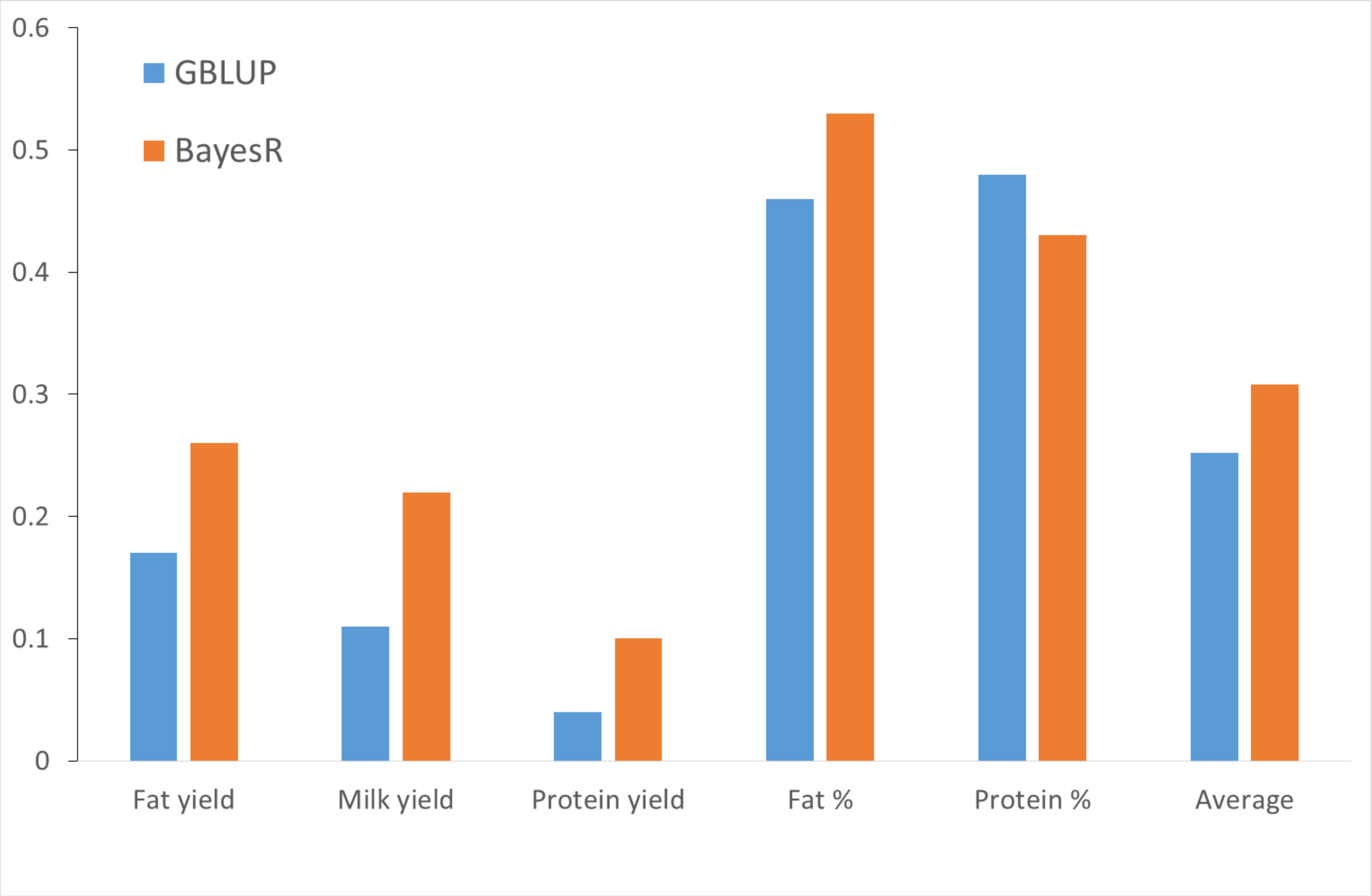
Intro

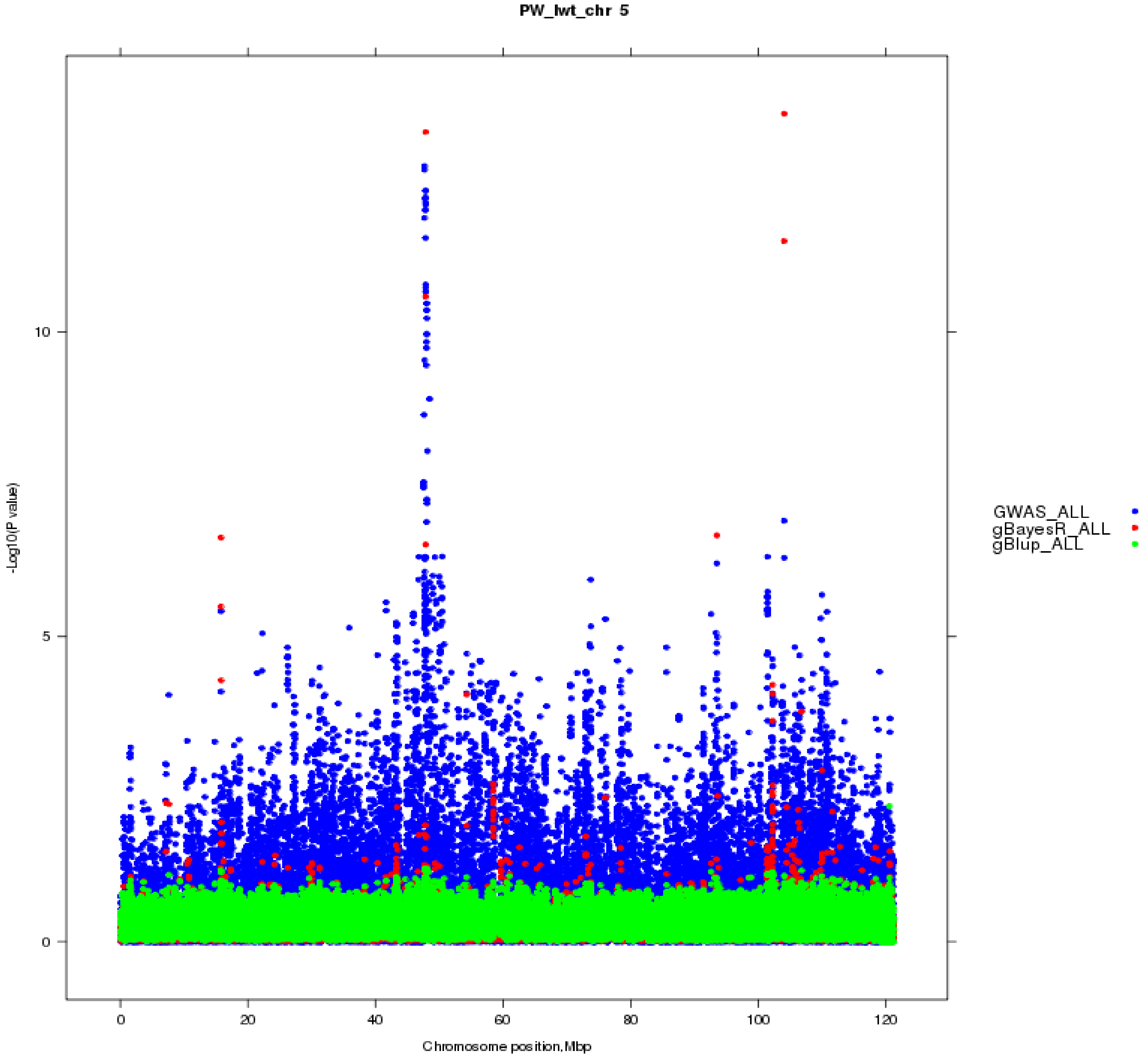
Non linear method (i.e. Bayesian method)

- Higher accuracy
- More robust
- Benefits from multiple breeds
- Benefits from genome sequence data
- Benefits from biological knowledge

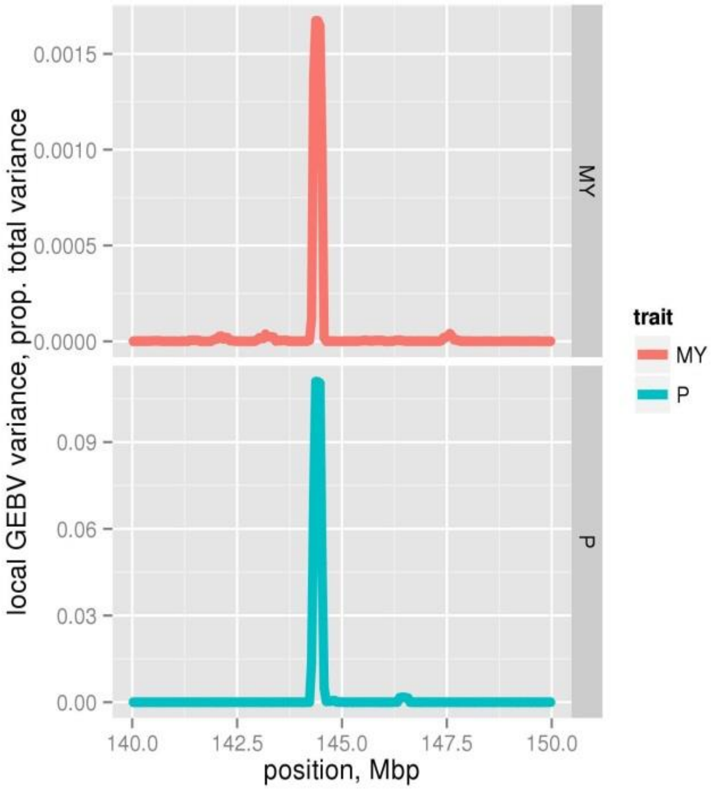
Combining one-step and non-linear method is research problem

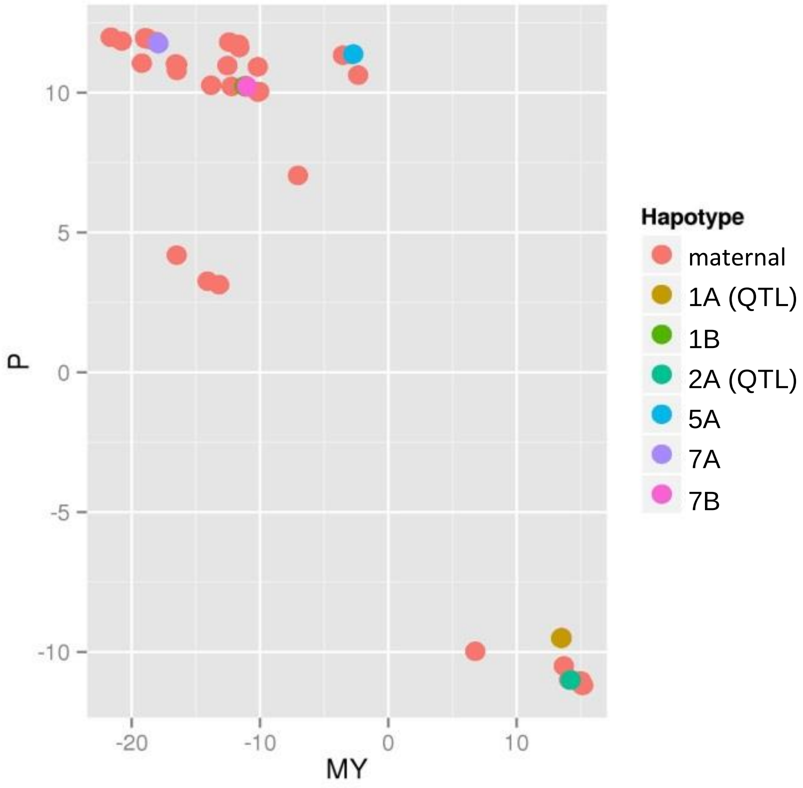
Accuracy r(DGV,DTD) in Aussie Red Bulls




PW_lwt_chr 5
-Log10(P value)
0
5
10
0
20
40
60
80
100
120
Chromosome position,Mbp
GWAS_ALL ●
gBayesR_ALL ●
gBlup_ALL ●

e.g. pleiotropic milk yield and phosphorus QTL on BTA1





	Effect	P-value	Prop. σ^2_P
Additional traits			
phosphorus conc.	**41.8**	**1.10×10^{-11}**	**0.107**
eSLC37A1	**0.160**	**3.55×10^{-18}**	**0.224**
Key production trait, milk yield			
milk yield – Holstein cows	**-37.6**	**2.19×10^{-3}**	**0.001**
milk yield – Holstein bulls	**-40.3**	**3.17×10^{-3}**	**0.003**
milk yield – Jersey cows	**-45.2**	**3.26×10^{-3}**	**0.002**

That is the allele that *increases* expression of SLC27A1 (an antiporter):

1. *Increases* phosphorus concentration
2. *Decreases* milk yield

Intro

Evaluating bulls → Evaluating SNPs

Interbull

Proposal

Short term

Combine BLUP SNP solutions for 50k SNP chip using SNP-MACE

Medium term

Identify sequence variants for inclusion in national evaluation

Short term
SNP-MACE

Within country i data can be modelled

y_i = other effects + $Z_i g_i + e_i$

where $V(g_i) = B_i$ diagonal

$V(e_i) = R_i$ diagonal

After absorbing other effects, can mimic equations by

$(Z_i' R_i^{-1} Z_i + B_i^{-1})\, g_i = Z_i' R_i^{-1}\, y_i$

Short term
SNP-MACE

Multiple country data

$y' = (y_i\ y_j),\ g' = (g_i\ g_j),\ e' = (e_i\ e_j)$

$$V(g) = B = \begin{array}{ll} (B_{ii} & B_{ij}) \\ (B_{ji} & B_{jj}) \end{array}$$

$$V(e) = R = \begin{array}{ll} (R_{ii} & 0\ \ ) \\ (0 & R_{jj}) \end{array}$$

$$(Z'R^{-1}Z + B^{-1})\ g = Z'R^{-1}\ y$$

Short term
SNP-MACE

Multiple country data

$(Z'R^{-1}Z + B^{-1})\ g = Z'R^{-1}\ y$

$$\begin{array}{llll} (Z_i'R_i^{-1}Z_i + B^{ii} & B^{ij} & &)\ g_i = Z_i'R_i^{-1}\ y_i \\ (& B^{ij} & Z_j'R_j^{-1}Zj_i + B^{jj} &)\ g_j = Z_j'R_j^{-1}\ y_j \end{array}$$

That is, we need the $Z_i'R_i^{-1}Z_i$, $Z_i'R_i^{-1}\ y_i$ and B_i from each country and the r_g between countries only.

Note, if we have $Z_i'R_i^{-1}Z_i$ and B_i , we can compute $Z_i'R_i^{-1}\ y_i$

That is, we only need the equations used by each country not the original data.

Short term
SNP-MACE

Fall back position

$Z_i'R_i^{-1}Z_i$ is a 50k x 50k matrix.
If we cant get it?

Use diagonal elements of $Z_i'R_i^{-1}Z_i$ and approximate off-diagonals by a sample of $Z_i'Z_i$
(Yang et al 2012)

And compute by $Z_i'R_i^{-1} y_i$
by
$(Z_i'R_i^{-1}Z_i + B_i^{-1})\ g_i = Z_i'R_i^{-1} y_i$

Short term
SNP-MACE

Complications

1. $C(e_i, e_j) \neq 0$ because some phenotypes used in both countries

   causes off-diagonal blocks in $Z'R^{-1}Z$

   can approximate by number of shared animals between countries

2. Different SNPs used in different countries

   Back solve from GEBVs to get equivalent SNP solutions for any SNP set

Short term
SNP-MACE

Integration with national evaluations

No one-step within country

Use Interbull SNP solutions instead of local ones

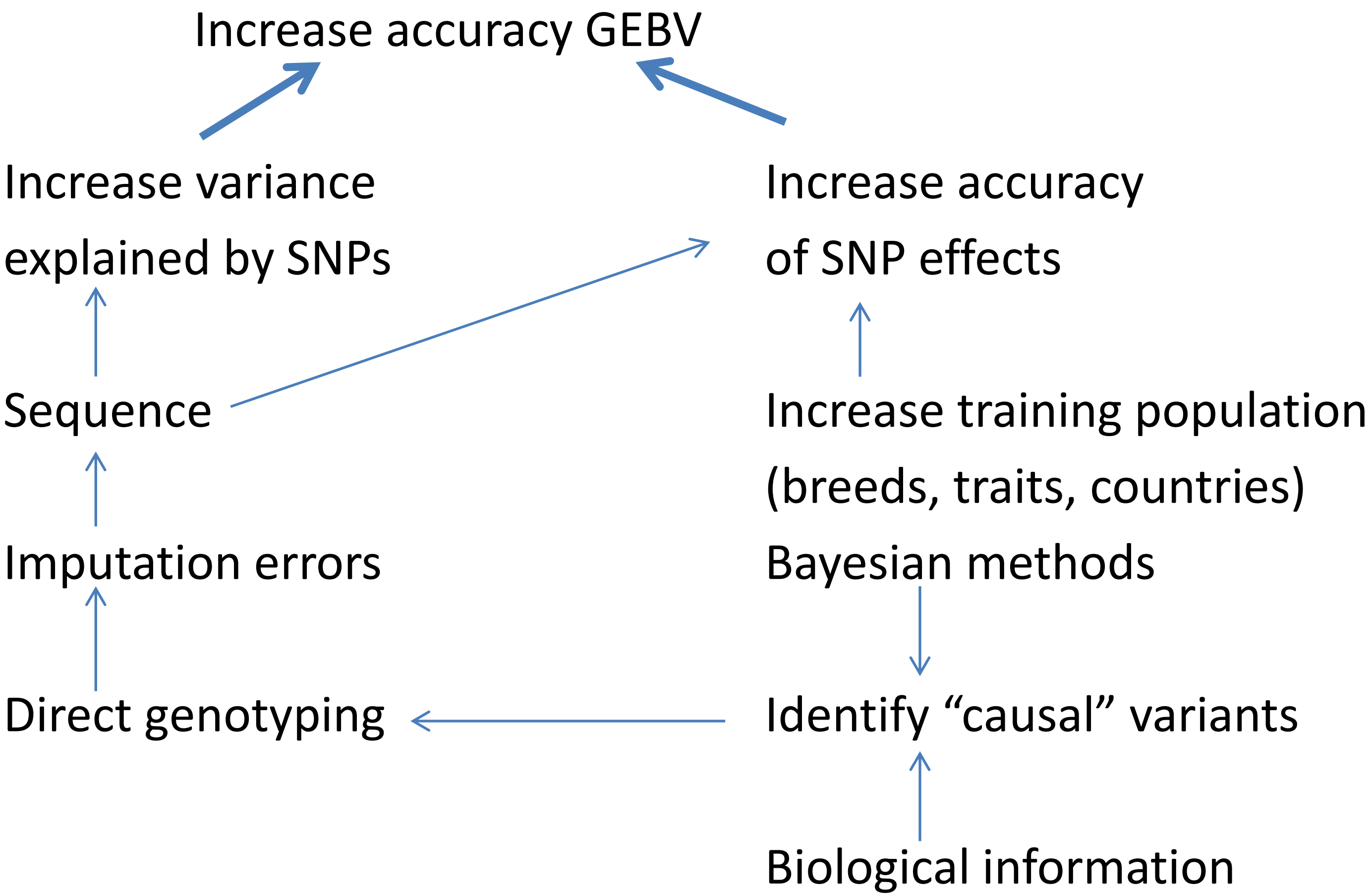
One step within country

Use one-step equations with SNP solutions and import Interbull solutions

OR

Treat Interbull SNP solutions as external data in local analysis

Medium term



Increase accuracy GEBV

Increase variance explained by SNPs

Increase accuracy of SNP effects

Sequence

Increase training population (breeds, traits, countries)

Imputation errors

Bayesian methods

Direct genotyping

Identify “causal” variants

Biological information

Medium term

Sequence variants + Bayesian methods + use of biological information

Benefits

- Increased accuracy
- Increased stability
- Multiple breed prediction
- Prediction of breed differences

Medium term

Sequence variants + Bayesian methods + use of biological information

How to do it?

Same data as BLUP

$Z_i'R_i^{-1}Z_i$ and $Z_i'R_i^{-1} y_i$

At sequence level

Impute effects

Countries impute genotypes

Countries genotype “causal” variants

Medium term 2

Estimate genetic correlation between countries

Estimate from SNP data instead of pedigree

→ less biased estimates

Uses same data as SNP-MACE

Proposal

Two analyses

1) Production run

   generates EBVs

   one step

   fast

   limited number of SNPs

   could use SNP variances or SNP effects estimated elsewhere

   could use BLUP

Proposal

Two analyses

2) Research run

Generates list of SNPs or SNP variances or SNP effects

Slow

Not necessarily one-step

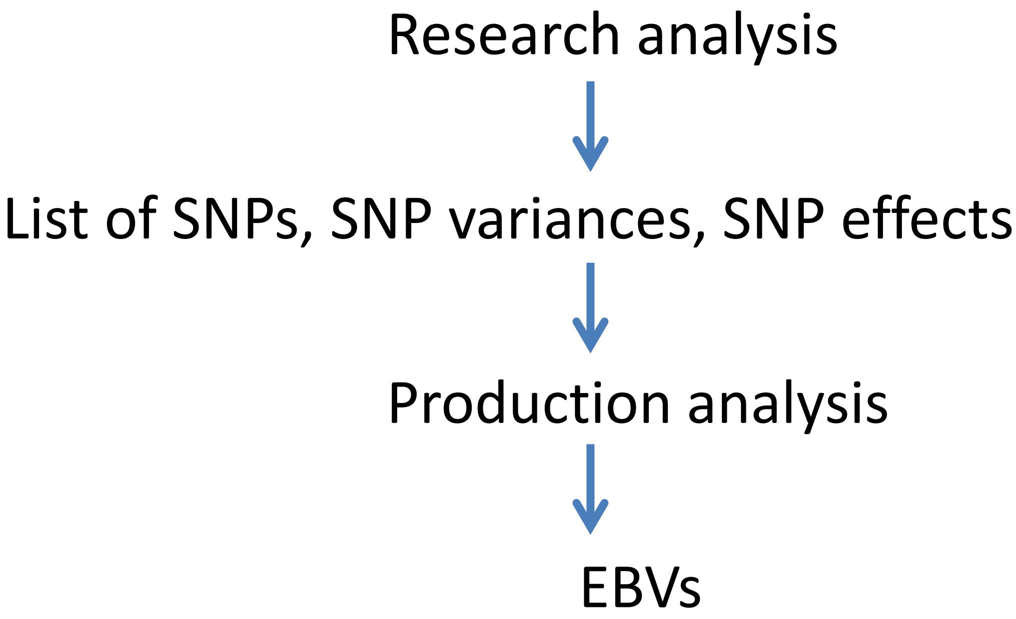
Large number of SNPs (WGS)

Non-linear method (EM plus MCMC)

First eliminate most SNPs from model

Second estimate remaining SNPs

Proposal



Research analysis

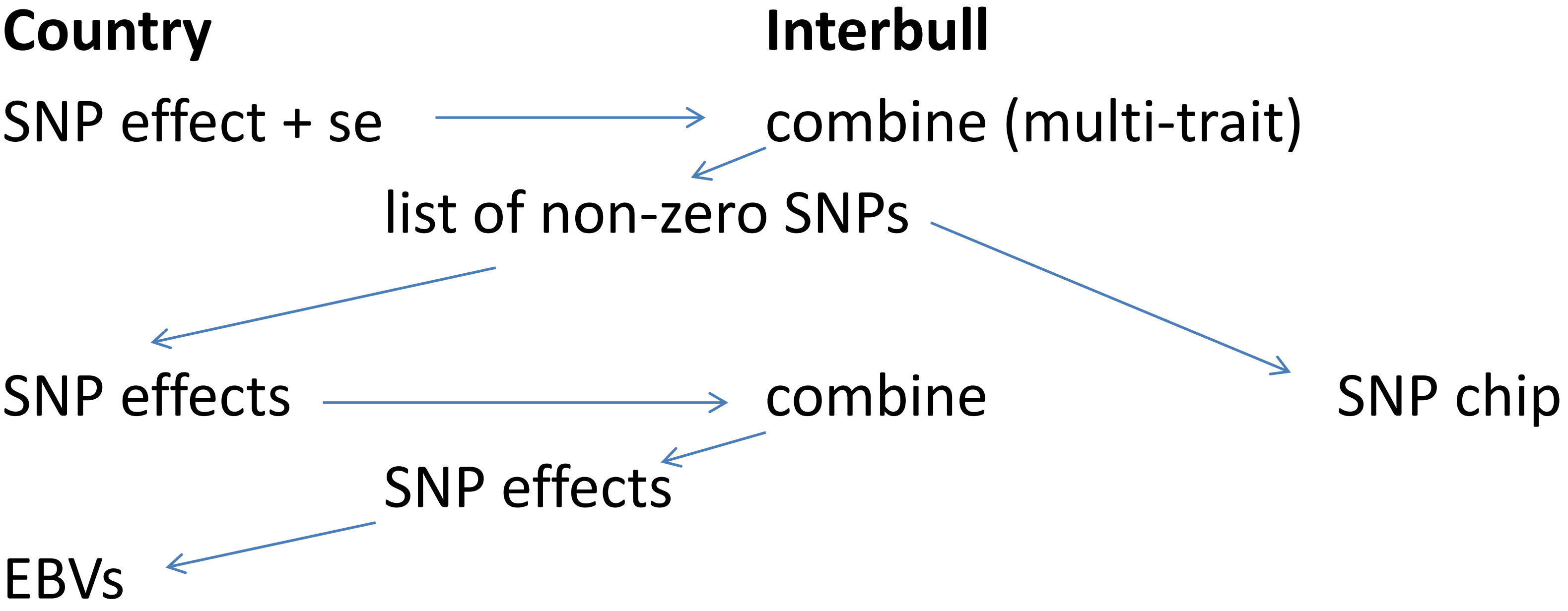
List of SNPs, SNP variances, SNP effects

Production analysis

EBVs

Proposal

Countries could collaborate through Interbull with one or both of these analyses



Country

Interbull

SNP effect + se

combine (multi-trait)

list of non-zero SNPs

SNP effects

combine

SNP chip

SNP effects

EBVs

Proposal

Short term

Combine BLUP SNP solutions for 50k SNP chip using SNP-MACE

Medium term

Identify sequence variants for inclusion in national evaluation

Estimate genetic correlations between countries

Benefits

Short term

More accurate GEBVs

Medium term

Common set of ‘causal SNPs’ are genotyped and used in GE

Multi-breed EBVs with increased accuracy

More robust EBVs

Increased understanding of our traits

Better genetic correlation estimates

Proposal

Actions

Decide to collaborate

Pilot project to implement SNP-MACE (short term)

Research project to find best methodology (medium term)