Meta-analysis of GWAS to estimate SNP effects and breeding values

Michael Goddard
University of Melbourne
Introduction

What is a meta-analysis?

Why do we need it?

How good is it?

Examples

Proposal for Interbull
What is a meta-analysis

Combining the results from >1 analysis rather than combining the raw data

Why do we need it?

To increase power

To increase robustness

Cant combine raw data
How good is it?

Very widely used to combine information from medical trials

Some information is lost
  Covariances between estimates
  OLS vs GLS
Example 1 - Multi-trait GWAS

Bolormaa et al (2017)

Combine single trait GWASs on correlated traits
Example 2- Multi-country GWAS of Stature

Bouwman et al (2018)

Combine GWASs on stature from different breeds and countries
58,000 bulls from 17 countries-breeds
163 lead variants with $p<5\times10^{-8}$
Explaining 13% of variance
Example 2 - Multi-country GWAS of Stature
Example 3- LD score regression in humans

Uses summary data ie chi-square from GWAS and LD r² to estimate h² (Bulik-Sullivan et al 2015)

Can be used to estimate \( r_g \)

Assess the importance of prior information (Finucane et al 2015)
  eg H3K4me3 sites enriched 2 fold for effect on 27 disease traits

Less accurate than raw data?
Example 4- Joint and conditional GWAS in humans

Yang et al (2012)
Simultaneously estimating the effect of many SNPs on a trait
Least squares equation for multiple regression
$X'Xb = X'y$

$b$ from published data
$X'X$ from reference sample
$X'Xb \rightarrow X'y$ for sample

Bayesian multiple regression (Zhu and Stephens 2016)
Example 5- Combining eQTL and GWAS summary statistics

Zhu et al 2017

Regression of expression on phenotype for individual SNP associations
Opportunities in cattle

Many countries and breeds have genotypes and phenotypes but raw data cannot be combined.
Opportunities in cattle

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
Opportunities in cattle

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
Interbull combines progeny test EBVs from different countries
→ more accurate EBVs which are comparable regardless of country of origin

Selection of bulls now largely on genomic EBVs

Lose information if you combine GEBVs from different countries

Better to combine SNP solutions
Interbull SNPMace

Lose information if you combine GEBVs from different countries

Options

Convert GEBVs from country A to country B as for progeny test EBVs
GEBVs are regressed back by $r_g$

Put genotype from country A through prediction equation of country B
Limited accuracy due to size of reference population in country B

Combine estimates of SNP effects, allowing for $r_g$, to get most accurate estimate of SNP effects in country B
Interbull SNPMace

Single country equations to estimate SNP effects (g)

\[(Z_1'Z_1 + \lambda I) g_1 = Z_1'y_1\]

Two countries

\[(Z_1'Z_1 + Z_2'Z_2 + \lambda I) g = Z_1'y_1 + Z_2'y_2\]

If individual countries provide \(g_i\) and \(Z_i'Z_i\) we can construct the multi-country BLUP and solve for \(g\)

Extensions: include \(r_g < 1\) between countries and weights for records
Extensions for SNPMace model

\[
\begin{pmatrix}
Z_1'Z_1 + G^{1} \\
\Psi_{12} + G^{12}
\end{pmatrix}
\begin{pmatrix}
\Psi_{12} + G^{12} \\
Z_2'Z_2 + G^2
\end{pmatrix}
\times
\begin{pmatrix}
\hat{g}_1 \\
\hat{g}_2
\end{pmatrix}
= \begin{pmatrix}
Z_1'y_1 \\
Z_2'y_2
\end{pmatrix}
\]

\[
G = \text{var}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}^{-1} = \begin{pmatrix}
\sigma_1^2 B_1 \\ r_{i2} \sqrt{\sigma_1^2 B_1 \times \sigma_2^2 B_2} \\
2 \sqrt{\sigma_1^2 B_1 \times \sigma_2^2 B_2} \\
\sigma_2^2 B_2
\end{pmatrix}^{-1}
\]

\[
B = \frac{1}{\sum_j 2p_i(1-p_i)} I
\]

(VanRaden 2008)
A general SNP Mace model

\[
\begin{bmatrix}
Z_i'Z_i + G^{ii} \\
\vdots \\
\Psi ii^+ + G^{ii^+} \\
\vdots \\
Z_{i^+}'Z_{i^+} + G^{i^+i^+} \\
\vdots \\
\end{bmatrix} 
\times 
\begin{bmatrix}
\hat{g}_i \\
\vdots \\
\hat{g}_{i^+} \\
\vdots \\
\end{bmatrix} 
= 
\begin{bmatrix}
Z_i'y_i \\
\vdots \\
Z_{i^+}'y_{i^+} \\
\vdots \\
\end{bmatrix}
\]
Interbull SNPMace Project

2018-2019

We will write software to do SNPMace and deliver it to Interbull

We will test method on Brown Swiss

  Interbull have individual records for all countries and can calculate GEBV
  We will use SNPMace to calculate SNP effects and hence GEBVs
  Compare the two sets of GEBVs

Based on 50k SNP genotypes
Interbull Data

- Z'R^{-1}Z matrices were calculated for six countries:

<table>
<thead>
<tr>
<th>Country</th>
<th>No bulls</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHE</td>
<td>1922</td>
</tr>
<tr>
<td>DEA</td>
<td>2578</td>
</tr>
<tr>
<td>FRA</td>
<td>171</td>
</tr>
<tr>
<td>ITA</td>
<td>1418</td>
</tr>
<tr>
<td>SVN</td>
<td>227</td>
</tr>
<tr>
<td>USA</td>
<td>796</td>
</tr>
</tbody>
</table>
Interbull SNPMace Project

Possible extension of project

Use all sequence variants and Bayesian method instead of BLUP.
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
Variance explained by SNPs and sequence (Iona Macleod)

Proportion of Total Genetic Variance Explained by SNP and Pedigree: BayesR (Mixed Hol & Jer)

- **% Genetic Var - SNP**
- **% Genetic Var - Ped**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Aust Bull &amp; Cow: Holstein &amp; Jersey</th>
<th>Danz Bulls Only: Holstein &amp; Jersey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperament</td>
<td>SEQ 50K</td>
<td>SEQ 50K</td>
</tr>
<tr>
<td>Stature</td>
<td>SEQ 800K</td>
<td>SEQ 800K</td>
</tr>
<tr>
<td>Milk Yield</td>
<td>SEQ 10K</td>
<td>SEQ 10K</td>
</tr>
<tr>
<td>Protein Yield</td>
<td>SEQ 50Kpruned</td>
<td>SEQ 50Kpruned</td>
</tr>
</tbody>
</table>

- 0.04 ~ 0.6 million
- 0.4 ~ 600,000
- 4 ~ 40,000
- 400,000 ~ 1 million
- 10,000 ~ 40,000
Accuracy $r(DGV,DTD)$ in Aussie Red Bulls

(Iona MacLeod)
Meat Traits:

GBLUP Accuracy - Merino x Border Leicester

Prediction Accuracy

- cc fat
- cem d
- im f
- pem d
- sf5
- pwt

50K
50K+Top Seq (2)
Wool Traits:
Prediction Accuracy in Merinos

- Breech Wrinkle
- Clean Fleece Wt
- Fibre Diameter

Prediction Accuracy:
- 50K
- BayesR 50K+Top Seq
- BayesRC 50K+Top Seq

![Bar Chart](chart.png)
BayesR vs BLUP (BTA11)

O = BayesR
O = GBLUP

Chromosome Position (Mb)
Cattle stature (Aniek Bouwman, Ben Hayes et al)

<table>
<thead>
<tr>
<th>Annotation class</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>intergenic_variant</td>
<td>83</td>
</tr>
<tr>
<td>upstream_gene_variant</td>
<td>11</td>
</tr>
<tr>
<td>5_prime_UTR_variant</td>
<td>1</td>
</tr>
<tr>
<td>intron_variant</td>
<td>55</td>
</tr>
<tr>
<td>missense_variant</td>
<td>5</td>
</tr>
<tr>
<td>downstream_gene_variant</td>
<td>8</td>
</tr>
<tr>
<td>ChiP-SEQ peaks*</td>
<td>8</td>
</tr>
<tr>
<td>WBC eQTL</td>
<td>10</td>
</tr>
</tbody>
</table>
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
EBVs → SNP effects
Conclusions

Meta-analysis could be used more to collaborate between countries

Interbull project is an example
Acknowledgements

Thank you to the Brown Swiss community for access to their data