Large Scale Genomic Evaluation in Dairy Cattle

Zengting Liu
F. Seefried, F. Reinhardt, and R. Reents
vit w.V., Heideweg 1, 27283 Verden, Germany
OUTLINE

■ Introduction
■ An example of genomic model (German Holsteins)
■ Topics relevant for large scale genomic evaluation
  1) Deregressing (inter)national EBVs
  2) SNP effect estimation
  3) Approximation of reliabilities of direct genomic values
■ Discussion
  ▪ Combining genomic with conventional information
■ Summary
Introduction: genetic evaluation models

- Conventional genetic evaluation in dairy cattle
  - Using phenotypic data and pedigree
  - Henderson’s mixed model (BLUP) methodology
  - Indirect setup of $A^{-1}$ and iteration on data techniques
    - 292 million test-day records from 16 million cows (German Holstein)
    - Total # equations: 370 millions for each of milk, fat, protein and SCS
  - Reliabilities of (multi-trait) EBV reasonably accurately approximated
  - Very successful for breeding, BUT
    - reliable EBV available late
    - low reliability for cows or pedigree index

- Genomic model contrasting the animal model
  - Small $n$ large $p$ problem of the genomic SNP model
  - Realised genomic ($G$) vs. expected relationship ($A$)
  - No algorithm yet for indirect setup of $G^{-1}$
  - Direct inversion of $G$ becoming increasingly infeasible
A BLUP SNP genomic model (German Holsteins)

- A BLUP SNP model for bulls (no cows in training set)

\[ q_i = \mu + v_i + \sum_{j=1}^{p} z_{ij} u_j + e_i \]

- Phenotypic record: bull deregressed EBV (similar to DYD)

\[ \text{var}(q_i) = \sigma_a^2 + \sigma_e^2 / \varphi_i \] where \( \varphi_i \) is effective # daughters of bull

- Deregressed EBV (DPRF) easier to obtain than DYD from international conventional evaluation MACE

- Residual polygenic effect with trait-specific variance
  - Validation showing candidates’ GEBV with too high variance
  - SNP markers may not explain all genetic variation
  - SNP effects depend too much on pedigree (Habier et al. 2007)
  - SNP effects less biased and more persistent (Solberg et al. 2009)
  - Polygenic effect also in French QTL model (Guillaume et al. 2008)
1) Deregressed EBV for genomic evaluation

- Genomic evaluation using national genotypes and phenotypes
  - Genotyped foreign calves may have sires without daughters in Germany
- EuroGenomics / North-American / IGenoP projects
  - Multiple country SNP model is a better approach (sharing genotypes?)
  - Use international MACE EBV for genomic evaluation
- Best possible choice for dependent variable
  - EBV should be avoided, due to double counting phenotypic info
  - DYD preferred, but not available for all traits / countries
  - Sub-optimal deregression on an animal by animal basis
    - DPRF = (EBV – PA)/$R_{dau}^2 + \mu$
    - Deregressing MACE EBV using full pedigree
      - Loop over pedigree sorted by birth years
      - Keep EBV constant for bulls with daughters
      - Iterative process until deregressed EBV converged
1) The MACE EBV deregression method

- MACE EBV and equivalent effective daughter contribution (EDC)
  - Required for all bulls on a given country scale
  - Even for bulls without local daughters

- Calculation of equivalent EDC for every bull
  - Using multi-trait EDC method (Liu et al. 2004)
  - National EDCs from all countries $\phi_i (i = 1, \ldots, 27)$
  - Genetic correlations between all country pairs
  - Sire variances for all countries
  - National heritability values

- Same software as for deregressing national EBV
### 1) Results: deregressed with original EBV

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>5685</td>
<td>95.8</td>
</tr>
<tr>
<td>1991</td>
<td>5809</td>
<td>95.7</td>
</tr>
<tr>
<td>1992</td>
<td>6156</td>
<td>95.8</td>
</tr>
<tr>
<td>1993</td>
<td>5937</td>
<td>95.1</td>
</tr>
<tr>
<td>1994</td>
<td>6206</td>
<td>96.0</td>
</tr>
<tr>
<td>1995</td>
<td>6438</td>
<td>95.7</td>
</tr>
<tr>
<td>1996</td>
<td>6661</td>
<td>96.0</td>
</tr>
<tr>
<td>1997</td>
<td>6816</td>
<td>95.5</td>
</tr>
<tr>
<td>1998</td>
<td>6459</td>
<td>95.7</td>
</tr>
<tr>
<td>1999</td>
<td>6156</td>
<td>95.5</td>
</tr>
<tr>
<td>2000</td>
<td>5940</td>
<td>95.7</td>
</tr>
<tr>
<td>2001</td>
<td>5963</td>
<td>96.0</td>
</tr>
<tr>
<td>2002</td>
<td>5977</td>
<td>96.0</td>
</tr>
<tr>
<td>2003</td>
<td>6009</td>
<td>95.1</td>
</tr>
<tr>
<td>2004</td>
<td>4089</td>
<td>93.8</td>
</tr>
<tr>
<td>2005</td>
<td>387</td>
<td>90.2</td>
</tr>
</tbody>
</table>

All Holstein bulls included with EDC>0

Milk yield

06 September 2010
1) Discussion: MACE EBV deregression

- Deregression very fast and well converged (milk yield)
  - 114,003 Holstein bulls with daughters worldwide
  - 212,181 Holstein animals in pedigree
  - 4494 rounds reached convergence \((10^{-10})\) in 2.3 minutes

- Many more bulls considered than national data (110,000 vs 24,000)

- Deregressed MACE reasonably highly correlated with MACE proofs

- The lower correlations of MACE than German national data:
  - Lower reliabilities of daughter info on German scale than national German proofs (0.75 vs 0.93)

- Larger difference between proofs and deregressed proofs for MACE than German national data
  - Pedigree difference between deregression (sire+dam) and MACE Aug’09 evaluation (sire+MGS+MGD group)
  - Unofficial bulls were missing in MACE result files
2) SNP effect estimation: Data materials

- **Genotyped animals (June’10)**
  - 27,721 genotyped animals in total
  - 17,477 Holstein bulls (5,477 DEU + 12,000 EuroGenomics)
    - 50,516 ancestors for estimating residual polygenic effect (RPG)
    - 107 phantom parent groups of RPG
    - Representing 21.4 million cows

- **Phenotypes from April’10 conventional evaluation**
  - Deregressed MACE or national proofs for 44 traits
  - 24,405 bulls with daughters in national evaluation
  - 114,003 bulls with daughters in international evaluation

- **Combining genomic and conventional evaluation**
  - 128,126 animals with (genomic) data
  - Including reference bulls, bulls with phenotype only, and candidates
  - 236,873 animals in pedigree and 330 phantom parent groups
2) SNP effect estimation: Computing resources

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>No. training bulls</th>
<th>No. animals in pedigree</th>
<th>CPU (snds) per round</th>
</tr>
</thead>
<tbody>
<tr>
<td>German national Aug’09</td>
<td>4339</td>
<td>24,478</td>
<td>16</td>
</tr>
<tr>
<td>German national Jan’10</td>
<td>5025</td>
<td>27,041</td>
<td>17</td>
</tr>
<tr>
<td>EuroGenomics Apr’10</td>
<td>17,429</td>
<td>50,385</td>
<td>99</td>
</tr>
<tr>
<td>EuroGenomics Jun’10</td>
<td>17,477</td>
<td>50,516</td>
<td>103</td>
</tr>
</tbody>
</table>

- CPU time increased linearly with no. of training bulls
- RAM usage increased also nearly linearly (2.9 Gb)
- Estimating residual polygenic effects required little CPU
- Convergence criteria: DGV > RPG > SNP effects
- Model with higher residual polygenic variance converged better
- SNP model feasible for very large reference population
2) Convergence of genomic model estimates

Residual polygenic variance: ~0%

SNP effects
Residual polygenic effect
DGV (sum of SNP)
GEBV = DGV + RPG
2) Convergence of genomic model estimates

Residual polygenic variance: 1%

SNP effects
Residual polygenic effect
DGV (sum of SNP)
GEBV = DGV + RPG

Milk yield
2) Reduced impact of genetic relationship on direct genomic values (training bulls)

Regression on sire EBV (y = $b_0 + b_1 \cdot \text{EBV}_{\text{sire}}$)

11,987 reference bulls with 580 genotyped sires

Trait: milk yield
2) Reduced impact of genetic relationship on direct genomic values (validation bulls)

Regression on sire EBV ($y = b_0 + b_1 \cdot \text{EBV}_{\text{sire}}$)

- 1211 validation bulls with 111 genotyped sires

- Direct genomic values (DGV)
- Combined GEBV

Trait: milk yield
3) Reliability of DGV (ΣSNP): introduction

- Calculation of reliability of DGV estimates
  - Method 1: ONE single value for ALL genotyped animals
    - Realised genomic reliability obtained from validation
  - Method 2: inverting genomic relationship matrix $G$
    - animal specific reliability
    - By-product of G-matrix BLUP method

- Direct matrix inversion approach
  - Desirable properties
  - Overestimation problem mainly due to the assumption of all SNPs are in complete LD with QTLs & IBS = IBD
    - Correct level derived from validation study
  - Less feasible for large-scale genomic evaluation
    - Though special software for inverting large matrices (up to 50,000 animals)
  - Approximation needed
3) Data materials for reliability approximation

- German national genomic evaluation Jan 2010
  - 10,487 genotyped animals
    - 5025 Holstein bulls in reference population
    - 5344 genotyped Holstein animals as candidates

- Reliability values of estimated DGV for the candidates
  - Obtained by direct matrix inversion
  - Used as reference value (response variable)

- Prediction formulae for approximating the reliabilities
  - Calculating various statistics as predictor variables
  - Selecting the best subset regressions
  - Using $R^2$ value and MSE for model comparison
3) Reliability method: Predictor variables

- Genomic relationship of a CANDIDATE to reference animals
  - Average with all reference animals:
    \[
    \bar{g}_i = \frac{\sum_{j=1}^{n} g_{ij}}{n}
    \]
  - Squared average value:
    \[
    \bar{g}_i^2
    \]
  - Maximum relationship value:
    \[
    g_{i}^{\text{max}} = \max(g_{i1}, \ldots, g_{in})
    \]
  - Sum of squared relationships:
    \[
    g_i^2 = \frac{\sum_{j=1}^{n} g_{ij}^2}{n}
    \]

- Reliability of individual reference animal
  - Daughter reliability:
    \[
    pg_i^2 = \left(\sum_{j=1}^{n} g_{ij}^2 \ast \frac{g_{jj}(1 - REL^{DAU}_j)}{\lambda}\right)/n
    \]
  - Genomic reliability:
    \[
    gg_i^2 = \left(\sum_{j=1}^{n} g_{ij}^2 \ast \frac{g_{jj}(1 - REL^G_j)}{\lambda}\right)/n
    \]
  - Genomic – daughter reliability:
    \[
    dg_i^2 = \left(\sum_{j=1}^{n} g_{ij}^2 \ast \frac{g_{jj}(1 - (REL^G_j - REL^{DAU}_j))}{\lambda}\right)/n
    \]

In total, 12 predictor variables studied, 5 insignificant
3) Results: correlation with genomic reliability

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>Correlation with candidates’ genomic reliabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average genomic relationship $\bar{g}_i$</td>
<td>.66</td>
</tr>
<tr>
<td>Squared relationship value $\bar{g}_i^2$</td>
<td>.64</td>
</tr>
<tr>
<td>Maximum relationship value $g_i^{\text{max}}$</td>
<td>.61</td>
</tr>
<tr>
<td>Sum of squared relationship $g_i^2$</td>
<td>.72</td>
</tr>
<tr>
<td>Daughter reliability REF $pg_i^2$</td>
<td>.71</td>
</tr>
<tr>
<td>Genomic reliability REF $gg_i^2$</td>
<td>.71</td>
</tr>
<tr>
<td>Genomic-daughter rel. REF $dg_i^2$</td>
<td>.72</td>
</tr>
</tbody>
</table>
3) Results: Optimal subset regressions

- Subset regressions
  - All combinations of 12 variables considered
    - R² value increased with # fitted variables
    - Balance R² and # variables

- Optimal subset regression for reliability prediction
  - Average genomic relationship to all training bulls \( \bar{g}_i = (\sum_{j=1}^{n} g_{ij}) / n \)
  - Squared average relationship \( \bar{g}_i^2 \)
  - Maximum genomic relationship \( g_{ij}^{\text{max}} = \max(g_{i1}, \ldots, g_{in}) \)
  - Sum of squared relationship \( g_i^2 = (\sum_{j=1}^{n} g_{ij}^2) / n \)

\[
REL_i = b_0 + b_1 \cdot \bar{g}_i + b_2 \cdot \bar{g}_i^2 + b_3 \cdot g_{ij}^{\text{max}} + b_4 \cdot g_i^2
\]

- Reliability of individual training bulls no longer important
- Optimal subset regressions CONSISTENT for all traits
- All 4 variables highly correlated, except \( g_{ij}^{\text{max}} \)
3) Results: Average genomic relationship $\bar{g}$.

Genomic reliability value and average genomic relationship

\[
\text{r}_{\text{elgn}} = 0.2672 + 6.0174 \times \text{avgGrel} - 27.126 \times \text{sqavgG}
\]
3) Results: Maximum genomic relationship $g_{ij}^{\text{max}}$

Genomic reliability value and maximum genomic relationship

\[ \text{relnm} = 0.4385 + 0.2123 \times \text{maxGrel} \]

- Sire not genotyped
- Sire/fullsib genotyped

Non-return rate cow

German national reference population

06 September 2010
3) Discussion: Approximating DGV reliability

- A method developed for approximating DGV reliabilities
  - Using genomic relationship of candidate to ALL training animals
  - 4 predictor variables selected

- Reasonably high goodness of fit achieved

- Approximated reliabilities to be adjusted to REALISED reliability level via validation study

- New derivation is needed, if reference population changes significantly in size and structure:
  - German national vs. EuroGenomics reference population
    - Size change: 5025 vs 17,054 Holstein bulls (Jan’2010)
    - Structural change in terms of genomic relationship:
      - Lower average genomic relationship for DEU candidates \( g_i \downarrow \)
      - Many more candidates with genotyped sire \( g_{i_{\text{max}}} \uparrow \)
Discussion: Combining genomic with conventional phenotypic information

* Turned out to be more difficult than initially assumed
  * Non-independent information sources DGV and EBV / PI

* Three alternatives for combining DGV and EBV
  1. Selection index with DGV as a new info source
     * Assume non-zero residual covariance
     * Limitation of inverting very large G matrix
     * Problem in modelling international original phenotypes in case of joint genomic reference populations
  3. Reasonable simplifications:
     * Use DYD/DPRF instead of ORIGINAL phenotypic records to remove all other effects (e.g. HYS, p.e. effects)
     * Use bulls rather than cows (Germany: 25,000 bulls vs. 17 mln cows)
Combining genomic with conventional info

- Three alternatives for combining DGV and EBV
  3. BLUP ‘pseudo-record’ method (Ducrocq & Liu, 2009)
    - Transforming genomic info into equivalent phenotypic records
    - Takes advantage of existing efficient BLUP software
    - Correct for genomic pre-selection bias (Patry & Ducrocq, 2009; Liu et al. 2009)
    - Automatic propagation of genomic info to non-genotyped relatives
    - The genomic LHS and RHS terms may be well approximated
    - Approximation of DGV reliabilities for candidates (Liu et al. 2010)
  4. Two correlated trait approach (Mäntysaari & Stranden, 2010)
    - Correlation of DGV and EBV for validation bulls
    - Problem of properly handling training animals

- Implementations of the methods may need fine tuning via validation study
Summary

- Large-scale conventional evaluation formed solid basis for genomic selection
- The genomic model is highly efficient
- Deregressed EBV as dependent variable for (inter)national genomic evaluation in dairy cattle
- Fitting residual polygenic effect may be necessary
  - To avoid too high variance of direct genomic values
  - To reduce pedigree impact on direct genomic values
- Approximation of DGV reliability is necessary with ever increasing number of genotyped animals
- Optimal large-scale combination of genomic and conventional information is important for comparable GEBV and EBV
- More R&D is needed for fine tuning
Acknowledgements

- FUGATO & FBF (GenoTrack)
- EuroGenomics Consortium
- Colleagues of Interbull Technical Committee
- Colleagues of Interbull Genomics Task Force
THANK YOU!

vit

IT-Solutions for Animal Production