Bioinformatic challenges in the analysis of complex traits

Joanna Szyda
Tomasz Suchocki, Magdalena Frąszczak, Magda Mielczarek

Wrocław University of Life and Environmental Sciences, Institute of Genetics, Biostatistics Group
1. Data

2. Estimation of genomic breeding values

3. Gene networks

4. Phenologs

5. GWAS

6. Rare variants

7. Comparison of populations

8. Conclusions
- **Polish Holstein-Friesian dairy cattle**
  - black-white & red-white males & females

- **Complete records on:**
  - pedigree, phenotypes, environmental factors

- **5 362 genotyped animals**
• **Single Nucleotide Polymorphisms SNP**
  \{ AA, AB, BB \}

• **Illumina Bovine50K SNP chip**
  54001 SNP v1
  54609 SNP v2

• **Data editing**
  call rate $\geq 90\%$
  MAF $\geq 0.01$

• **46 267 SNPs**

Data: genotypes
• Phenotypes routinely recorded in Poland

• Complex mode of inheritance: major genes + polygenes

• Quantitative traits & score traits

• **Pseudophenotypes** for bulls
  calculated based on daughters’ data
  sum of additive effects of all genes
  breeding values
• 3 production traits $h^2 \approx 0.30$
• 1 udder health (somatic cell score) $h^2 \approx 0.30$
• 21 type & conformation $h^2 \in [0.10, 0.54]$
• 4 fertility $h^2 \approx 0.02$
predict genetic quality of young animals

• **Direct Genomic Value (DGV)**

• **Sum of additive effects of SNPs**

• **Estimates of individual SNP effects → GWAS**

• **Technical challenge:**
  - large data → model dimensions
  - real data → errors
  - repeated evaluation → robust model
DGV: reference animals

- Reference group
- Pseudophenotypes → Daughters
- SNP genotypes → Own

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
\[ y = \mu + Zq + e \]

- **y** pseudophenotype \([1 \times 2\,761]\)
- **q** SNP effects \( \sim N\left(0, I \frac{\sigma_a^2}{46\,267}\right) \) \([1 \times 46\,267]\)
- **Z** \( \in \{-1, 0, 1\} \) \([2\,761 \times 46\,267]\)
- **e** residual \( \sim N\left(0, D\sigma_e^2\right) \) \([2\,761 \times 2\,761]\)
SNP effects on fat yield

SNP effects on somatic cell score
test group

pseudophenotypes $\rightarrow$ daughters

SNP genotypes $\rightarrow$ own

\[ DGV_i = Z_i \hat{q}_i \]
Gene network: motivation

Identify (all) genes underlying a complex trait

- Basic data
  - SNP position → Illumina + manually corr.
  - SNP pairwise LD → PLINK
  - Gene position → Ensemble rel.68

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
**Gene network: gene effect estimation**

**SNP effect estimates (q)**

- Genomic location + pairwise LD ($r^2$)

**Gene effect estimates (g)**

\[
g = \frac{\sum \hat{q}_i}{\sigma_g}
\]

\[
\sigma^2_g = \sum \sigma^2_{qi} + 2 \sum \sum \sigma_{qij}
\]

\[
\sigma^2_g = n\sigma^2_q + 2 \sum_i \sum_{j>i} r_{ij}^2 \sigma^2_q
\]
Gene network: gene selection

estimates for 4,345 genes

\[ g \sim N(0,1) \rightarrow P \text{ value} \]

\[ P \text{ value} < 0.20 \]

C14H8orf33, AGO2, RHPN1, GML, EBS1, MAF1, MAPK15, DGAT1, LY6D

milk yield
Gene network: network construction

Bisogenet, Martin et al. 2010 BMC Bioinformatics

retrieve functional information

326 KEGG pathways
Kobas, Xie et al. 2011 Nucleid Acids Research

2 289 GO terms
Bisogenet, Martin et al. 2010 BMC Bioinformatics
Gene network: network validation

- SNP effect estimation
- Gene effect estimation
- Gene selection
- Network construction
- Functional information

Phenotype permutation

Data DGV estimation Gene networks Phenologs GWAS Rare variants Population comparison
Gene network: network validation

- SNP effect estimation
- Gene effect estimation
- Gene selection
- Network construction
- Functional information

KEGG

GO

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
Gene network: testing

Odds Ratio for KEGG/GO

\[ H_0 : P(O) = P(P) \quad H_1 : P(O) \neq P(P) \]

\[ \ln(OR) = \ln \left( \frac{\frac{C_O}{(N_O - C_O)}}{\frac{C_P}{(N_P - C_P)}} \right) \]

- original data
- permuted data (pooled)

\[ \sim N \left( 0, \sigma^2_{\ln(OR)} \right) \rightarrow \sim N(0,1) \rightarrow \text{Bonferroni} \]
Gene network: results GO

- regulation of translation → P<0.00001
- down regulation of translation involved in gene silencing by miRNA → P<0.00001
- RNA-mediated gene silencing → P<0.00001
- cytoplasmic mRNA processing body → P<0.00001
- RNA-induced silencing complex → P=0.00060
- double-stranded RNA binding → P=0.00333
- down reg. of translational initiation → P=0.01487
- pre-miRNA processing → P=0.03088
- hemidesmosome assembly → P=0.03630
Gene network: results KEGG

- **Galactose metabolism** (30 genes) → $P=0.01357$
- **Pentose phosphate** (26) → $P=0.03223$
- **Fructose and mannose metabolism** (36) → $P=0.03223$
- **Measles** → $P=0.04278$
- **Dilated cardiomyopathy** → $P=0.05933$
- **p53 signaling pathway** → $P=0.09567$
- **Hypertrophic cardiomyopathy** → $P=0.09567$
Gene network: results KEGG

- Galactose metabolism (30 genes) → $P=0.01357$
- Pentose phosphate (26) → $P=0.03223$
- Fructose and mannose metabolism (26) → $P=0.03223$
- Immunogenesis → bacterial infection susceptibility
- Measles → $P=0.04278$
- Dilated cardiomyopathy → $P=0.05933$
- p53 signaling pathway → $P=0.09567$
- Hypertrophic cardiomyopathy → $P=0.09567$
Gene network: results KEGG

- Galactose metabolism (30 genes) → P=0.01357
- Pentose phosphate (26) → P=0.03223
- Fructose and mannose metabolism (36) → P=0.03223
- Measles → P=0.04278
- Dilated cardiomyopathy → P=0.05933
- p53 signaling pathway → P=0.09567
- Hypertrophic cardiomyopathy → P=0.09567
express similarity between traits via physiological information

Prediction of gene–phenotype associations in humans, mice, and plants using phenologs

John O Woods, Ulf Martin Singh-Blom, Jon M Laurent, Kriston L McGary and Edward M Marcotte

Woods et al. BMC Bioinformatics 2013, 14:203
http://www.biomedcentral.com/1471-2105/14/203

RESEARCH ARTICLE Open Access
Phenolog: gene networks

Gene networks

Bisogenet
Martin et al. 2010 BMC Bioinformatics

GSLA
Zhou et al. 2013 Bioinformatics

fat yield
milk yield
protein yield
somatic cell
stature

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
**Gene Matrix**

<table>
<thead>
<tr>
<th>gene</th>
<th>milk</th>
<th>fat</th>
<th>protein</th>
<th>cell</th>
<th>stature</th>
</tr>
</thead>
<tbody>
<tr>
<td>KCNAB1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DLG4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SUMO3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAV1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

...
Do different methods choose the same SNPs?

\[ P < 0.001 \]

- Single SNP and polygenic effect
  \[ \rightarrow \text{Wald test with Bonferroni} \]

- Nonparametric CAR score
  \[ \rightarrow \text{based on empirical CV} \]

- DGV estimation model
  \[ \rightarrow \text{Wald test without Bonferroni} \]
GWAS: results

SNP effects significant for protein yield

M2: Single SNP and polygenic effect
M3: Nonparametric CAR score
M4: Genomic selection model
Do rare SNP variants explain significant part of genetic variation?

- MAF > 0.01
- Call rate > 0.95
- 46,267 SNP

- Call rate > 0.95
- 53,867 SNP
Rare variants: results

SNP effects on protein yield
- common SNPs
- common & rare SNPs

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
Comparison of populations: motivation

If / where two dairy cattle populations differ?

2,243 bulls  
2,294 bulls

the same
• SNPs
• traits
• statistical model

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
Comparison of populations: LD

Data  DGV estimation  Gene networks  Phenologs  GWAS  Rare variants  Population comparison
Comparison of populations: correlation

- Milk yield: $r=0.15$
- Fat yield: $r=0.19$
- Protein yield: $r=0.19$

Data, DGV estimation, Gene networks, Phenologs, GWAS, Rare variants, Population comparison
1. More dense SNP chips
   • E.g. 777 K

2. NGS
   • 1 000 bull genomes project

3. New phenotypes
   • Eco: methane emission, feed efficiency

4. Other livestock species
   • poultry, horses, pigs


• **genes of medium / small effects difficult to capture**
  → ... but important for trait variation
  → variety of analysis models needed

• **account for nonadditive effects dominance / epistasis**
  → ... low statistical power
  → large data sets needed

complex are traits determined by many genes
Thank you ...