Accounting for read depth in the analysis of genotyping-by-sequencing data

Ken Dodds, John McEwan, Timothy Bilton, Rudi Brauning, Rayna Anderson, Tracey Van Stijn, Theodor Kristjánsson, Shannon Clarke

AgResearch, Invermay, New Zealand
Genotyping-by-sequencing (GBS)

- Alternative genotyping technology
- Methods here apply to any sequencing-based genotypes
- Reduced representational methods cost-effective
  - Restriction enzyme digests
  - Untargeted regions
  - Barcode to multiplex lanes (48 – 384)
  - No-oligo design/purchase required
  - Reference sequence optional
  - Based on Elshire method (‘GBS’)

GBS SNPs

- True heterozygote (AB)
- Observation probabilities (random model)

<table>
<thead>
<tr>
<th># reads ( (k) )</th>
<th>Only A’s ( k )</th>
<th>A’s and B’s ( k )</th>
<th>Only B’s ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>0.75</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>0.0625</td>
<td>0.875</td>
<td>0.0625</td>
</tr>
<tr>
<td>( \infty )</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Inferred genotype

- AA*
- AB*
- BB*

* denotes inferred genotype
Genomic relatedness theory

• $P(\text{AA}^*|\text{AB}) = K$
• $P(\text{AB}^*|\text{AB}) = 1 - 2K$

• $P(\text{AA}^*) = P(\text{AA}) + P(\text{AB})K$

• $P(\text{AA}) = p^2 + p(1 - p)F$

• $P(\text{AA}, \text{AA}) = \text{function of relatedness measures}$
• …

* denotes inferred genotype

For random sampling: $K = 1/2^k$
$k = \text{depth}$

$F = \text{inbreeding}$
Genomic relatedness theory

- **M**: genotypes \((0/1/2)\) (individuals x SNPs)
- Centre using \(P\) (\(j^{th}\) column = \(2p_j\))
  - \(Z = M - P\)
- \(G_1 = \frac{zz'}{2\Sigma p_j(1-p_j)}\)
- Numerator for 1 SNP: \(z_{ii'} = (x_i - 2p)(x_{i'} - 2p)\)
  - \(x_i\) is the marker score \((0, 1, 2)\) for \(g_i^*\)
- \(E(z_{ii'}) = 2p(1-p)2\theta\)
- \(E(z_{ii}) = 2p(1-p)(1 + F_i + 2K_i - 2F_iK_i)\)
Genomic relatedness theory: Remarks

• \( G_1 = \frac{zz'}{2\sum p_j(1-p_j)} \)

• \( E(z_{ii'}) = 2p(1-p)2\theta \)

• \( E(z_{ii}) = 2p(1-p)(1 + F_i + 2K_i - 2F_iK_i) \)

• \( G_1 \) divisor gives the correct expected value
  • Assumes \( p \) is known
Genomic relatedness theory: Remarks

- \( G_1 = \frac{zz'}{2 \sum p_j (1 - p_j)} \)
- \( E(z_{ii'}) = 2p(1 - p)2\theta \)
- \( E(z_{ii}) = 2p(1 - p)(1 + F_i + 2K_i - 2F_i K_i) \)

- The expected value for different individuals does not depend on \( K \)
- The expected value for self-relatedness depends on \( K \)
  - No information on \( F \) when \( k = 1 \)
    (discard)
  - Need to take depth into account (\( k > 1 \))
Genomic relatedness theory

- Missing values
  - Usual method is ‘naïve imputation’ – use $2p_j$
    - Biased downwards
  - Our method: only use SNPs non-missing (for both individuals)

- Kinship using GBS with Depth adjustment (KGD)

- Dodds et al (2015) BMC Genomics

- R code on github
Example: parentage

- 2203 Atlantic Salmon (pedigree recorded)
- 122 full-sib families, (122 dams, 66 sires, 1177 progeny)
- Genotyped: all sires, 119 dams and 94 full-sib families
- Filter to exclude duplicated regions (HW-.05)
- 30,923 SNPs; mean SNP depth was 7.9 (All)
- 24,899 SNPs; mean SNP depth was 3.3 (HW-.05)

### Mean relatedness estimates

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Number of SNPs</th>
<th>Identity</th>
<th>Full-sibs</th>
<th>Parent-Offspring</th>
<th>Non-sib Offspring</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>30,923</td>
<td>0.739</td>
<td>0.375</td>
<td>0.382</td>
<td>0.109</td>
</tr>
<tr>
<td>HW-.05</td>
<td>24,899</td>
<td>1.014</td>
<td>0.454</td>
<td>0.461</td>
<td>0.014</td>
</tr>
</tbody>
</table>
Example: GRM

Dairy Sheep
Also 15k chip genotyped

Data courtesy of Suzanne Rowe, AgResearch

Off-diagonal comparisons

Diagonal comparisons

Depth
- ≤ 0.3
- 0.3 - 1
- 1 - 2
- >2

Chip

GBS

y = -0.0991 (SE 0.02) + 1.19 (SE 0.02) x

n = 1539
r = 0.939

Spring Sheep - Milk Co. - New Zealand
Example: Genomic Prediction

- GBLUP Breeding values
- Milk (kg)
GBS: Linkage Disequilibrium

- Use Likelihood approach
  - Observed genotype combination given LD, read depth

- Simulations
  - True Values:
    - $D' = 0.6$
    - $r^2 = 0.36$
  - Standard method results in strong bias
  - Depth-adjusted method improves bias but is more variable
  - The two methods were similar with high read depth
LD Estimation: Application

- Genome assembly improvement
- White clover
- Allotetraploid
- Full-sib family
- Marker pairs in “backcross” configuration

Data courtesy of Andrew Griffiths, AgResearch
Conclusions

• Low depth GBS …
  • Analysis methods to accommodate data type
    • Allelic sampling, rather than genotypes

• Method for unbiased relatedness estimation
  • Many genetic analyses can be based on relatedness estimates

• Method for Linkage disequilibrium

• Enables low depth GBS to be used
Acknowledgements

Genomics for Production & Security in a Biological Economy

Suzanne Rowe (dairy sheep)
Andrew Griffiths (clover)