A hidden Markov model to estimate inbreeding from whole genome sequence data

Tom Druet & Mathieu Gautier
Unit of Animal Genomics, GIGA-R, University of Liège, Belgium
Centre de Biologie pour la Gestion des Populations, INRA, France
Introduction

• Controlling inbreeding in livestock species or in small populations
  – Recessive defects, inbreeding depression, etc.

• Genomic data
  – Observation of realized inbreeding
  – Pedigree sometimes unavailable
Genomic inbreeding $F$

- Estimation with genomic relationship matrix (GRM)
  - Reference population
  - Independent SNPs
  - Global estimate

- Runs of homozygosity (ROH)
  - Parameter definitions
  - Allele frequencies not used
  - Inappropriate for low-fold sequencing
Hidden Markov models

- Models the genome as a mosaic of IBD (inbred) and non-IBD segments (e.g., Leutenegger, 2003 - AJHG)

10020110102111100200202021211012110210110120101210011
Hidden Markov models

• Models the genome as a mosaic of IBD (inbred) and non-IBD segments (e.g., Leutenegger, 2003 - AJHG)
Emission probabilities

- Probability of genotype given IBD status (emission prob.):

<table>
<thead>
<tr>
<th></th>
<th>IBD</th>
<th>Non-IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_iA_i$</td>
<td>$p_i$</td>
<td>$p_i^2$</td>
</tr>
<tr>
<td>$A_iA_j$</td>
<td>$\epsilon$</td>
<td>$2p_ip_j$</td>
</tr>
</tbody>
</table>
Transition probabilities

- Absence of coancestry change is $e^{-\alpha}$ ($\alpha$ is the transition rate: recombination rate & time to common ancestor)
- Prob. new coancestry is IBD is $F$
- Prob. New coancestry is non-IBD equals $(1-F)$
### Transition probabilities

- **Transition matrix:**

<table>
<thead>
<tr>
<th></th>
<th>IBD</th>
<th>Non-IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBD</td>
<td>((1-e^{-\alpha})(1-F))</td>
<td>((1-e^{-\alpha})(1-F))</td>
</tr>
<tr>
<td>Non-IBD</td>
<td>((1-e^{-\alpha})F)</td>
<td>((1-e^{-\alpha})F)</td>
</tr>
</tbody>
</table>
## Transition probabilities

- Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>IBD</th>
<th>Non-IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBD</td>
<td>$e^{-\alpha}$</td>
<td>$(1-e^{-\alpha})(1-F)$</td>
</tr>
<tr>
<td>Non-IBD</td>
<td>$(1-e^{-\alpha})F$</td>
<td>$e^{-\alpha}$</td>
</tr>
</tbody>
</table>
Transition probabilities

- Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>IBD</th>
<th>Non-IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBD</td>
<td>$e^{-\alpha} + (1-e^{-\alpha})F$</td>
<td>$(1-e^{-\alpha})(1-F)$</td>
</tr>
<tr>
<td>Non-IBD</td>
<td>$(1-e^{-\alpha})F$</td>
<td>$e^{-\alpha} + (1-e^{-\alpha})(1-F)$</td>
</tr>
</tbody>
</table>
Extension to WGS data

• Replace genotypes in emission probabilities:
  – Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF):

\[
P(\text{Data} \mid \text{IBD}) = p_i P(A_iA_i \mid \text{Data}) + p_j P(A_jA_j \mid \text{Data}) + \epsilon P(A_iA_j \mid \text{Data})
\]
Extension to WGS data

- Replace genotypes in emission probabilities:
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  - Use allele counts (allele depth – AD)

\[
P(\text{AD} \mid \text{IBD}) = p_i P(\text{AD} \mid A_iA_i) + p_j P(\text{AD} \mid A_jA_j)
\]

ε included
Extension to WGS data

• Replace genotypes in emission probabilities:
  – Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  – Use allele counts (allele depth – AD)

• Recent implementations:
  – BCFtools / RoH (Narasimhan et al. – Bionformatics, 2016)
  – ngsF-HMM (Viera et al. – Bionformatics, 2016)
Limitation

• Assumes a single inbreeding event (one ancestor)
  – Still a single reference population

• In livestock species, complex inbreeding
  – Many common ancestors over many generations
  – Variable Ne over time (including bottlenecks)
Mixture of inbreeding classes

• Mixture of several IBD and nonIBD with different age (G)

• Emission probabilities unchanged

• Transition probabilities same principle
  – Each distribution with its own mixing proportions
Mixture of inbreeding classes

• Mixture of several IBD and nonIBD with different age (G)

• Emission probabilities unchanged

• Transition probabilities same principle
  – Each distribution with its own mixing proportions
Testing with simulations

- One distribution (1 age), 500 individuals, medians
Estimated F ~ Simulated F

- Simulated F = 0.05 and G = 64
Two simulated distributions

- Simulated Age, $G_1 = 16$ & $G_2 = 256$
Two simulated distributions

- Mixture of 10 predefined classes (9 IBD, 1 nonIBD)
Summary of simulations

- Simulations with varying age, number of distributions, type of markers, low-fold sequencing data, errors

- Assessing with estimated age, mixing (1 dist.), global F, local F, population and individual estimates, estimating K

- Better when younger F, larger F, more markers, higher MAF, higher cover, large age differences
Belgian Blue cattle (634 bulls)

Proportion inbreeding per age class

Total F

LD
50K
HD
ROH HD 1Mb+
ROH HD 100kb+
WGS data (high cover @114x)

- Sire x MGS mating: expected 25% at G3
WGS data (high cover @114x)

- Sire x MGS mating: expected 25% at G3

<table>
<thead>
<tr>
<th>Chr</th>
<th>Length (Mb)</th>
<th>#het snps</th>
<th>#snps</th>
<th>Prop. het</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>92.385886</td>
<td>23</td>
<td>192567</td>
<td>1.2e-4</td>
</tr>
<tr>
<td>1</td>
<td>51.469735</td>
<td>0</td>
<td>117044</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>46.047682</td>
<td>1</td>
<td>107278</td>
<td>9.3e-6</td>
</tr>
<tr>
<td>16</td>
<td>44.281690</td>
<td>0</td>
<td>81934</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>34.592319</td>
<td>13</td>
<td>80042</td>
<td>1.6e-4</td>
</tr>
<tr>
<td>4</td>
<td>33.943960</td>
<td>4</td>
<td>84630</td>
<td>4.7e-5</td>
</tr>
<tr>
<td>4</td>
<td>32.406205</td>
<td>0</td>
<td>64784</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>30.317150</td>
<td>6</td>
<td>70982</td>
<td>8.4e-5</td>
</tr>
<tr>
<td>10</td>
<td>27.445232</td>
<td>2</td>
<td>62643</td>
<td>3.2e-5</td>
</tr>
<tr>
<td>23</td>
<td>26.648470</td>
<td>1</td>
<td>74953</td>
<td>1.3e-5</td>
</tr>
</tbody>
</table>
BBB WGS (@10-15x)

- Longest IBD segments for one sire

<table>
<thead>
<tr>
<th>Chr</th>
<th>Lenght (HD)</th>
<th>#Het</th>
<th>#SNPs</th>
<th>Length (WGS geno)</th>
<th>#Het</th>
<th>#SNPs</th>
<th>Prop. Het</th>
<th>Lenght (Gen. Lik)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>94.6</td>
<td>2</td>
<td>23298</td>
<td>84.6</td>
<td>375</td>
<td>182480</td>
<td>0.0025</td>
<td>94.6</td>
</tr>
<tr>
<td>22</td>
<td>46.4</td>
<td>1</td>
<td>11834</td>
<td>34.1</td>
<td>82</td>
<td>69465</td>
<td>0.0012</td>
<td>45.2</td>
</tr>
<tr>
<td>13</td>
<td>34.0</td>
<td>0</td>
<td>7031</td>
<td>31.3</td>
<td>141</td>
<td>59879</td>
<td>0.0023</td>
<td>34.1</td>
</tr>
<tr>
<td>20</td>
<td>20.6</td>
<td>0</td>
<td>5418</td>
<td>20.5</td>
<td>127</td>
<td>48748</td>
<td>0.0026</td>
<td>20.7</td>
</tr>
<tr>
<td>8</td>
<td>16.2</td>
<td>0</td>
<td>3331</td>
<td>9.3</td>
<td>41</td>
<td>19566</td>
<td>0.0021</td>
<td>16.2</td>
</tr>
</tbody>
</table>

BovineHD  WGS called genotypes  WGS likelihoods
BBB WGS (@10-15x)

- Repartition in IBD classes (geno vs gen. likelihoods)
Whole Genome Sequence

- 50 sequenced Belgian Blue sires

Inbreeding for 50 WGS sires
Conclusions

• The model uses all the information
  – Sequence of genotypes, allele frequencies, error rates
• The model classifies inbreeding in different age classes
  – Better than just one (open perspectives)
• The model estimates local and global inbreeding
• The model can work with genotyping arrays and sequence data
  – With different allelic spectra