Using Genomic Relationship Likelihoods to perform trio parentage assignment

Feb. 13, 2018
By: Kim E. Grashei
Jørgen Ødegård & Theo Meuwissen

Norwegian University of Life Sciences
Topics

• Genomic Relationship Likelihood (GRL)
• Iterative training on the evaluation dataset
• Simulation experiment
  • Comparison with Colony2 (likelihood-based)
• Using GRL assignments to train an exclusion-based method
Genomic Relationship Likelihood (GRL)

• Trio assignment (child with unknown parents, candidate mother and candidate father)
• Uses genomic relationships
  • VanRaden’s (2008)* Method 1
    • Separate nominator and denominator

*VanRaden P.M., Efficient methods to compute genomic predictions, J Dairy Sci. 91 (2008) 4414-4423
Genomic relationships - Logic

- $r_{ij}$ is the genomic relationship between animals $i$ and $j$
  - $E(r_{o,o} | \text{parents}) = 1 + 0.5r_{f,m}$
  - $E(r_{o,f} | \text{parents}) = 0.5(r_{f,f} + r_{f,m})$
  - $E(r_{o,m} | \text{parents}) = 0.5(r_{m,m} + r_{f,m})$

- Relationship residuals:
  - $e_{o,o} = r_{o,o} - E(r_{o,o} | \text{parents})$
  - $e_{o,f} = r_{o,f} - E(r_{o,f} | \text{parents})$
  - $e_{o,m} = r_{o,m} - E(r_{o,m} | \text{parents})$

- Residual relationships are assumed $\sim$ normally distributed
  - $e \sim N(\mu, \Sigma)$
- Iteratively assigned trios used as training data to estimate $\mu$ and $\Sigma$

Assume: “f” and “m” are the true parents of “o”

Normal assumption for residuals:

- $e \sim N(\mu, \Sigma)$

Graphs showing density distribution with error percentages of 0%, 1%, and 3%.
Assigning trios

- $GRL = -\frac{1}{2}(e - \mu)'\Sigma^{-1}(e - \mu)$
  - Proportional to the log of a multivariate normal density function
  - The value should be within the upper 99% interval of true parent trios
  - Threshold estimated using an iteration algorithm (next slide)
- $\Delta GRL = GRL_1 - GRL_2$
  - $GRL_1$ and $GRL_2$ are the most- and second-most likely trios for an offspring
  - The likelihood of the best trio should be at least 1,000 times higher than the second-best trio
  - $\Delta GRL \geq \ln(1000)$
- Only parent candidates with relationships to the offspring $> 0.25$ were considered
  - Reduces the number of tested trios
Training

• **Step 1, allele-dropping**: Mate random individuals in the evaluation dataset
  • Estimate $\mu$ and $\Sigma$
  • The $GRL$-threshold is chosen to be the cut-off value for the lower 1% of true trios in the simulation

• **Step 2, assignment iteration**
  • Use latest estimates of $\mu$ and $\Sigma$ to assign trios in the evaluation dataset
  • Re-estimate $\mu$ and $\Sigma$ using the assigned trios

• **Step 3**
  • Return to step 2 until the number of assigned trios starts to decrease

• **Step 4**
  • Discard the last iteration
  • Parameters from the second-last iteration are used as final estimates
Simulation

• QMSim* used for simulation of genotypes

• ~54k SNP markers

• 1,000 offspring with parents and 1,000 offspring without parents present

• Added genotype error and call rate variance

• Repeated 50 times

GRL – simulated data

![Bar chart showing count of correct, incorrect, and no-assign cases for available and missing parents.](chart.png)
Colony2 – simulated data

Available parents
Correct Incorrect No-assign

Missing parents
Incorrect No-assign

Count
0 100 200 300 400 500 600 700 800
Using GRL to train Exclusion-based method

- $E \sim Bin(n, p)$
- $E$: number of exclusions for a trio
- $n$: number of genotype calls for the trio
- $p$: median exclusion ratio of GRL-assigned trios
(GRL-trained) Exclusion method – simulated data
Conclusions

- GRL works well for assigning trios
- The training procedure can be used on an evaluation dataset containing true (but unknown) trios
  - Have to assume a $\Delta GRL$ threshold
- Parameters for exclusion-based methods can successfully be estimated using the GRL assignments
  - It might be useful having a parentage assignment method that does not rely on traditional likelihood- or exclusion-based methodology
Thank you for listening!