Progeny Array Statistical Genetics Methods

Vince Buffalo

December 15, 2015

1 Parentage Inference Methods

Parentage inference methods are inspired by those from Meagher and Thompson (1986), which are reviewed in Marshall, Slate, Kruuk, and Pemberton (1998). The major difference in our model is the incorporation of an error model that allows for different error rates for heterozygous and homozygous genotypes. We'll step through Meagher and Thompson's original model before deriving the model with error.

1.1 Original Model (without Genotyping Error)

Meagher and Thompson's basic model is to evaluate the relatedness of three individuals, M, F, and O, using a likelihood calculated from the joint probability of these individuals' genotypes. Here, M, F, and O represent alleged mother, alleged father, and offspring, respectively. For each loci, we calculate the probability of these individuals' genotypes g_m, g_f, and g_o (respectively) under two pedigree models: that M and F are the parents of O, and that M, F, and O are all unrelated individuals. Following Meagher and Thompson's notation, these two relationships are denoted QQ and UU. Another possible relationship between these individuals is that neither M or F is a parent of O; we denote this as QU (we'll skip including this possibility for now). Thus, for a single locus the joint probability of genotypes g_m, g_m, and g_o given that individuals M and F are O's parents is:

$$P(g_m, g_f, g_o|QQ) = T(g_o|g_m, g_f)P(g_m)P(g_f)$$

(from Meagher and Thompson 1986). Here, $T(g_o|g_m, g_f)$ is the Mendelian transmission matrix — essentially a conditional probability matrix. $P(g_m)$ and $P(g_f)$ are the probability of the observed genotypes for mother and father. Assuming random mating, these are just the Hardy-Weinberg genotype probabilities. For any genotype with alternate allele frequency p, the vector $\mathbf{h}$ encodes the genotype probabilities under Hardy-Weinberg:

$$\mathbf{h} = \begin{bmatrix} (1-p)^2 \\ 2p(1-p) \\ p^2 \end{bmatrix}$$

digraph structure gmp [texlbl=”g'_m”] gfp [texlbl=”g'_f”] gm [texlbl=”g_m”] gf [texlbl=”g_f”] go [texlbl=”g_o”] gop [texlbl=”g'_o”]
gmp -¿ gm [dir=back, color=”blue”] gfp -¿ gf [dir=back, color=”blue”] gm -¿ go [color=”orange”] gf -¿ go [color=”orange”] go -¿ gop [color=”blue”] rank=same; gmp; gfp rank=same; gm; gf

Figure 1: A directed graph model with hidden genotype variables g_m, g_f, and g_o, and observed genotypes g'_m, g'_f, and g'_o.

1.2 Genotype Likelihoods with Error

At this point, this model assumes g_m, g_f, and g_o are known with certainty. In reality, the true genotypes g_m, g_f, and g_o are hidden and only the genotypes with error g'_m, g'_f, and g'_o are observed. Graphically, this is depicted in Figure 2.

To model genotyping errors, we derive an expression for $P(g'_m, g'_f, g'_o)$ (given QQ) that marginalizes over the hidden genotype variables g_m, g_f, and g_o. So,

$$P(g'_m, g'_f, g'_o) = \sum_{g_m, g_f, g_o \in \Omega} P(g'_m, g'_f, g'_o | g_m, g_f, g_o) P(g_m, g_f, g_o)$$

The advantage of this approach is that we know through Mendelian transmission and Hardy-Weinberg that the component $P(g_m, g_f, g_o) = T(g_o|g_m, g_f)P(g_m)P(g_f)$, so,

$$P(g'_m, g'_f, g'_o) = \sum_{g_m, g_f, g_o \in \Omega} P(g'_m, g'_f, g'_o | g_m, g_f, g_o) T(g_o|g_m, g_f) P(g_m) P(g_f)$$

Now, $P(g'_m, g'_f, g'_o|g_m, g_f, g_o)$ seems complex, but each of these observed genotypes conditioned on the real genotype is independent of the others. For example, $g'_o \perp\!\!\!\perp g'_m \mid g_m$, meaning that knowing g'_m doesn't provide any information about the observed offspring's genotype g'_o given that we know g_m. This allows us to say $P(g'_m, g'_f, g'_o|g_m, g_f, g_o) = P(g'_m|g_m)P(g'_f|g_f)P(g'_o|g_o)$. So:

$$P(g'_m, g'_f, g'_o) = \sum_{g_m, g_f, g_o \in \Omega} P(g'_m|g_m) P(g'_f|g_f) P(g'_o|g_o) T(g_o|g_m, g_f) P(g_m) P(g_f)$$

Every component of this model is now tractable. Our error model enters through the stochastic transition matrix $\mathbf{E}$ for any true, latent genotype g to an observed genotype g':

$$\mathbf{E} = P(g'|g) = \begin{bmatrix} 1-e & e/2 & e/2 \\ \epsilon/2 & 1-\epsilon & \epsilon/2 \\ e/2 & e/2 & 1-e \end{bmatrix}$$

Where e is the homozygous error rate and ϵ is the heterozygous error rate. This error model is similar to models in which the probability of error is distributed uniformly over the two erroneous genotypes (Sobel, Papp, and Lange, 2002; Lincoln and Lander, 1992).

Extending this to all loci is trivial; we assume independence across loci, so for all loci we sum the log probability. Allowing locus l's joint probability to be written as $P(g'_{m,l}, g'_{f,l}, g'_{o,l})$, the probability of M and F being O's parents is:

$$P(G'_m, G'_f, G'_o|QQ) = \sum_{l \in L} \log P(g'_{m,l}, g'_{f,l}, g'_{o,l})$$

digraph parent$_{i}mputenode[fixedsize = shape]gmp[texlbl = $”$g'_m$”] gm [texlbl=”$g_m$”] go1 [texlbl=”$g_{o,1}$”] go2 [texlbl=”$g_{o,2}$”] go3 [texlbl=”$g_{o,3}$”] go4 [texlbl=”$g_{o,4}$”] go5 [texlbl=”$g_{o,5}$”] gop1 [texlbl=”$g'_{o,1}$”] gop2 [texlbl=”$g'_{o,2}$”] gop3 [texlbl=”$g'_{o,3}$”] gop4 [texlbl=”$g'_{o,4}$”] gop5 [texlbl=”$g'_{o,5}$”] gmp -¿ gm [dir=back, color=”blue”] gm -¿ go1 [color=”orange”] gm -¿ go2 [color=”orange”] gm -¿ go3 [color=”orange”] gm -¿ go4 [color=”orange”] gm -¿ go5 [color=”orange”] go1 -¿ gop1 [color=”blue”] go2 -¿ gop2 [color=”blue”] go3 -¿ gop3 [color=”blue”] go4 -¿ gop4 [color=”blue”] go5 -¿ gop5 [color=”blue”]

Figure 2: A half-sib (with shared mother) with five offspring.

1.3 Parental Genotype Imputation

$$P(g_m, g'_{o,.}) = P(g_m)\prod_{i=1}^{n} P(g'_{o,i}|g_m)$$

$$P(g_m, g'_{o,.}) = P(g_m)\prod_{i=1}^{n}\sum_{g_o\in\Omega} P(g'_{o,i}, g_{o,i}|g_m)$$

$$P(g_m, g'_{o,.}) = P(g_m)\prod_{i=1}^{n}\sum_{g_o\in\Omega} P(g'_{o,i}|g_{o,i})P(g_{o,i}|g_m) \tag{1}$$

The form of this model is similar to a naive Bayesian model (Koller and Friedman 2009, p. 50), with an added layer of uncertainty from true offspring genotypes to observed offspring genotypes.

1.4 Parental Haplotype Inference

Adopting the design of (Browning and Browning, 2009), we build a duo leveled HMM (H^2) for GBS data with high error rate. Our hidden states are the alleles on three haplotypes:

References

Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Stephen E Lincoln and Eric S Lander. Systematic detection of errors in genetic linkage data. *Genomics*, 14(3):604–610, 1992.

T.C. Marshall, J. Slate, L.E.B. Kruuk, and J.M. Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology*, 7(5):639–655, 1998.

Thomas R Meagher and Elizabeth Thompson. The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, 29(1):87–106, 1986.

Eric Sobel, Jeanette C Papp, and Kenneth Lange. Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, 70(2):496–508, 2002.