Notes on PPGtk

Paul Blischak

E-mail: blischak.4@osu.edu

Contents

1 Allele frequency estimation

1.1 The likelihood

The likelihood of an individuals' read data given the population allele frequency can be computed by summing over the possible genotypes:

$$\mathcal{L}_i(p) = P(r_i|p) = \sum_{a=0}^{m_i} P(r_i|a)P(a|p), \tag{S1}$$

where $P(r_i|a)$ is the genotype likelihood for genotype $a = 0, \ldots, m_i$ (e.g., calculated using GATK), and

$$P(a|p) = \binom{m_i}{a} p^a (1-p)^{m_i - a}.$$

For multiple samples, we take the product of the individual likelihoods:

$$\mathcal{L}(p) = \prod_i \mathcal{L}_i(p) = \prod_i \left(\sum_{a=0}^{m_i} P(r_i|a)P(a|p) \right). \tag{S2}$$

Taking the natural log gives us the log likelihood of the population allele frequency at a single site:

$$\ell(p) = \log \mathcal{L}(p) = \sum_i \log \left(\sum_{a=0}^{m_i} P(r_i|a)P(a|p) \right). \tag{S3}$$

1.2 Metropolis-Hastings algorithm

$$P(p) \sim \text{beta}(\alpha = 0.5, \beta = 0.5). \tag{S4}$$

$$P(p|r) \propto P(r|p)P(p) = \left(\sum_a P(r|a)P(a|p) \right) P(p) \tag{S5}$$

$$\alpha = \min \left\{ 1, \frac{P(r|p^*)P(p^*)}{P(r|p)P(p)} \right\} \tag{S6}$$

2 Inbreeding

We introduce another parameter, ϕ, that is related to the inbreeding coefficient (F) through the following equation:

$$F = \frac{1}{1+\phi} \tag{S7}$$

2.1 The likelihood

$$\mathcal{L}_i(p,\phi) = P(r_i|p,\phi) = \sum_{a=0}^{m_i} P(r_i|a)P(a|p,\phi) \tag{S8}$$

where $P(r_i|a)$ is the genotype likelihood for genotype $a = 0, \ldots, m_i$ (e.g., calculated using GATK), and

$$P(a|p,\phi) = \binom{m_i}{a} \frac{\mathcal{B}(a + p\phi, m_i - a + (1-p)\phi)}{\mathcal{B}(p\phi, (1-p)\phi)},$$

which is the probability density function for the beta-binomial distribution with $n = m_i$, $k = a$, $\alpha = p\phi$, and $\beta = (1-p)\phi$. Here, $\mathcal{B}()$ is the beta function:

$$\mathcal{B}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x\,.$$

2.2 Metropolis-Hastings algorithm

3 Beta mixture model

$$\begin{aligned}
P(g = a|p,\phi) &\sim \text{beta-binomial}(n = m_i, k = a, \alpha = p\phi, \beta = (1-p)\phi) \\
P(p|\gamma,\pi,\theta_1,\theta_2) &\sim \gamma P(p|\pi\theta_1,(1-\pi)\theta_1) + (1-\gamma)P(p|\pi\theta_2,(1-\pi)\theta_2), \\
P(\pi) &\sim \text{beta}(0.5, 0.5), \\
P(\gamma) &\sim \text{beta}(1, 1), \\
P(\phi) &\sim \text{gamma}(2, 0.1), \\
P(\theta_1) &\sim \text{gamma}(2, 0.1), \\
P(\theta_2) &\sim \text{gamma}(2, 0.1).
\end{aligned}$$

3.1 The likelihood

3.2 Metropolis-Hastings algorithm

4 Population admixture model

4.1 The likelihood

4.2 Metropolis-Hastings algorithm