# 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

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

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

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

## 1.3 Genotype estimation

$$P(g_i = a|r_i) = \frac{P(r_i|g_i = a)P(g_i = a|\hat{p})}{\sum_{j=0}^{m_i} P(r_i|g_i = j)P(g_i = j|\hat{p})} \tag{S7}$$

# 2 Disequilirium

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

$$F = \frac{1}{1+\phi} \tag{S8}$$

## 2.1 Derivative of the log beta-binomial function

The beta function is defined as: $\mathcal{B}(\alpha, \beta) = \int_0^1 t^{(\alpha-1)}(1-t)^{(\beta-1)}\mathrm{d}t$.

$$\alpha(p) = p\phi, \text{ and } \beta(p) = (1-p)\phi\,. \tag{S9}$$

$$\frac{\partial}{\partial p}\alpha(p) = \phi\,, \text{ and } \frac{\partial}{\partial p}\beta(p) = -\phi\,. \tag{S10}$$

$$\frac{\partial}{\partial p}\mathcal{B}(\alpha(p), \beta(p)) = \int_0^1 \frac{\partial}{\partial p}\left\{t^{(\alpha(p)-1)}(1-t)^{(\beta(p)-1)}\right\}\mathrm{d}t \tag{S11}$$

## 2.2 The likelihood

$$\mathcal{L}_i(p, \phi) = P(r_i|p, \phi) = \sum_{a=0}^{m_i} P(r_i|a)P(a|p, \phi) \tag{S12}$$

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:

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

## 2.3 Metropolis-Hastings algorithm

# 3 Allopolyploid mixture model

$$P(g_i = a|p_1, p_2, m_{i,1}, m_{i,2}) = \sum_{j=0}^{a}\sum_{r=0}^{a-j} P(g_{i,1} = j|p_1, m_{i,1})P(g_{i,2} = r|p_2, m_{i,2})\,. \tag{S13}$$

Here $m_{i,1} + m_{i,2} = m_i$ (e.g., 2:4 inheritance in an allopolyploid corresponds to $m_{i,1} = 2$ and $m_{i,2} = 4$) .

$$P(g_i = a|p_1, p_2, m_{i,1}, m_{i,2}) = \sum_{j=0}^{a} \left( \sum_{r=0}^{a-j} \left[ \binom{m_{i,1}}{j} p_1^j (1-p_1)^{m_{i,1}-j} \times \binom{m_{i,2}}{r} p_2^r (1-p_2)^{m_{i,2}-r} \right] \right). \quad \text{(S14)}$$

This is a special case of the Poisson-Binomial probability distribution where we have two different probabilities of success. The most general case would be where the probability of success for every allele is different.

$$\mathcal{L}(p_1, p_2) = \sum_{a=0}^{m_i} \mathcal{L}(g_i = a) P(g_i = a|p_1, p_2, m_{i,1}, m_{i,2}) \,. \quad \text{(S15)}$$

$$p_{\ell,k} \sim \text{Beta}(\alpha = \pi\theta_k, \beta = (1-\pi)\theta_k), \qquad \text{where } \ell = 1, \ldots, L, \text{ and } k = 1, 2. \quad \text{(S16)}$$

# 4 Beta mixture model

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

## 4.1 The likelihood

## 4.2 Metropolis-Hastings algorithm

# 5 Population admixture model

## 5.1 The likelihood

## 5.2 Metropolis-Hastings algorithm