Extending PPGtk

Paul Blischak

E-mail: blischak.4@osu.edu

Contents

1 Program organization

1.1 Model namespaces and classes

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

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

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

2.2 Metropolis-Hastings algorithm

3 Beta mixture model

3.1 The likelihood

3.2 Metropolis-Hastings algorithm

4 Population admixture model

4.1 The likelihood

4.2 Metropolis-Hastings algorithm