Appendix: Mutstats: An Ultra-fast Computational Method to Determine Clonal Status of Somatic Mutations

Dehua Bi, Subhajit Sengupta, Tianjian Zhou, and Yuan Ji

April 22, 2021

1 PyClone Model

PyClone performs Dirichlet Process (DP) clustering based on a hierarchical Bayes statistical model. Input to PyClone are the allelic counts from a set of N deeply sequenced mutations for a sample. Let N denotes the mutation, M the sample size, ϕ^n the cellular prevalence of mutation n across the M samples, the model is shown below:

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$$

$$H_0 = \text{Uniform}([0,1]^M)$$

$$H|\alpha, H_0 \sim \text{DP}(\alpha, H_0)$$

$$\alpha^n|H \sim H$$

$$\boldsymbol{\psi}_m^n|\boldsymbol{\pi}_m^n \sim \text{Categorical}(\boldsymbol{\pi}_m^n)$$

$$\boldsymbol{\psi}_m^n = (g_{m,N}^n, g_{m,R}^n, g_{m,V}^n)$$

either

$$b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m \sim \text{Binomial}(d_m^n, \psi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m))$$

or

$$s|a,b \sim \text{Gamma}(a_s, b_s)$$

$$b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m, s \sim \text{BetaBinomial}(d_m^n, \psi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m), s)$$

where

$$\psi(\boldsymbol{\psi}, \phi, t) = \frac{(1-t)c(g_N)}{Z}\mu(g_N) + \frac{t(1-\phi)c(g_R)}{Z}\mu(g_R) + \frac{t\phi c(g_V)}{Z}\mu(g_V)$$

$$Z = (1-t)c(g_N) + t(1-\phi)c(g_R) + t\phi c(g_V)$$

In this model, $a_\alpha = 1$, $b_\alpha = 10^{-3}$ for the DP concentration parameter α and $a_s = 1$, $b_s = 10^{-4}$ for the Beta Binomial precision parameter s. The Gamma distributions are parametrised in terms of the shape a and rate b.