Multiallelic calling model in bcftools (-m)

Petr Danecek, Stephan Schiffels, Richard Durbin

Version: April 22, 2016

Let x and y denote alleles. For simplicity of notation we work with SNPs, $x, y \in \{A, C, G, T\}$, but the method is identical for indels. We consider N samples. At a given site for sample i, let $Q^x_{i,1}, Q^x_{i,2}, \ldots$ be the quality scores of the reads covering the site. A simple estimate for the total allele frequency at the site is then simply:

$$f_x = \frac{\sum_i f^i_x}{N}, \tag{1}$$

with

$$f^i_x = \frac{\sum_k Q^x_{i,k}}{\sum_{k,y} Q^y_{i,k}}. \tag{2}$$

Note that equation 1 can be easily generalised to include a reference panel with N_{ref} samples and observed allele counts k^{ref}_x:

$$f_x = \frac{\sum_i f^i_x + k^{\mathrm{ref}}_x}{N + N_{\mathrm{ref}}}, \tag{3}$$

which allows the calling to benefit from prior allele frequency information on the site.

Now, given a particular allele set $S \subseteq \{A, C, G, T\}$, we introduce the relative frequencies

$$f_{x|S} = \frac{f_x}{\sum_{y \in S} f_y}. \tag{4}$$

We calculate the likelihood of observing the set of alleles S for each sample

$$L^i_S = \sum_{x,y \in S} f_{x|S}\, f_{y|S}\, G_i(xy), \tag{5}$$

where $G_i(xy)$ are the genotype likelihoods PL of i-th sample calculated by mpileup[1].

Finally, we have to give a prior probability for the allele set S. From basic population genetic theory, we know that the probability for a single mutation in a genealogy of N samples is given by θW_N, where W_N is the Watterson factor

$$W_N = \sum_{k=1}^{2N-1} \frac{1}{k}, \tag{6}$$

[1] $\mathrm{PL} = -10 * log_{10} P(\mathrm{data}|\mathrm{genotype})$

and θ is the scaled effective population size $\theta = 4N\mu$, which in humans is typically around $\theta = 0.001$.

We therefore impose a prior probability of

$$P(S) = (W_N\theta)^r, \tag{7}$$

where r is the number of non-reference alleles, i.e. the number of mutations needed to explain allele set S.

Given the prior probability $P(s)$, we can calculate the likelihood for all samples given allele set S as

$$L_S = (W_n\theta)^r \prod_i L_S^i. \tag{8}$$

Finally we select the most likely set of alleles $X \subseteq S$ so that

$$X = \arg\max_S L_S. \tag{9}$$

The site quality of variant sites is given by

$$\text{QUAL} = \frac{L_{\{ref\}}}{\sum_S L_S}, \tag{10}$$

where $\{ref\}$ denotes the reference allele, and the quality of non-variant sites

$$\text{QUAL} = 1 - \frac{L_{\{ref\}}}{\sum_S L_S}. \tag{11}$$

Assuming HWE, the most likely genotype $(xy)_i$ of i-th sample is

$$(xy)_i = \arg\max_{a,b \in X} L_X^i \tag{12}$$

and the corresponding genotype quality (the posterior genotype probability) is

$$\text{GQ} = \frac{L_X^i}{\sum_Y L_Y^i}. \tag{13}$$