Multiallelic calling model in bcftools (-m)

Petr Danecek, Stephan Schiffels, Richard Durbin

Version: June 13, 2014

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. Let's denote the number of samples N, the read depth of i-th sample N_i. In the pileup we observe the set of bases $S \subseteq \{A, C, G, T\}$, each base x is observed N^x times with the qualities $Q_1^x, \ldots, Q_{N^x}^x$. As a simple estimate of allele frequencies we take

$$f_x = \frac{\sum_k Q_k^x}{\sum_{k,y} Q_k^y}. \tag{1}$$

When calling jointly on multiple samples with varying coverage, lower-coverage samples would contribute less to the estimate. Therefore we calculate the frequencies f_x^i per-sample as indicated above and then calculate the site frequency as

$$f_x = \frac{\sum_i f_x^i}{N}. \tag{2}$$

Now, given a particular allele set S, we introduce frequencies

$$f_{x|S} = \frac{f_x}{\sum_{y \in S} f_y} \tag{3}$$

and define the likelihood for a sample i as

$$L_S^i = \sum_{x,y \in S} f_{x|S}\, f_{y|S}\, G^i(xy), \tag{4}$$

where $G^i(xy)$ are the genotype likelihoods PL calculated by mpileup[1]. Given the prior probability θ, the number of non-reference alleles r observed across all samples and using the Watterson factor W_n

$$W_N = \sum_{k=1}^{2N-1} \frac{1}{k}, \tag{5}$$

we calculate the overall likelihood for all samples given the allele set S as

$$L_S = (W_n \theta)^r \prod_i L_S^i. \tag{6}$$

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

$$X = \arg\max_S L_S. \tag{7}$$

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

The site quality of variant sites is given by

$$\text{QUAL} = \frac{L_{\{ref\}}}{\sum_S K_S} \tag{8}$$

and the quality of non-variant sites

$$\text{QUAL} = 1 - \frac{L_{\{ref\}}}{\sum_S K_S}. \tag{9}$$

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

$$(xy)^i = \arg\max_{a,b \in X} L_X^i \tag{10}$$

and the corresponding genotype quality is

$$\text{GQ} = \frac{L_X^i}{\sum_Y L_Y^i}. \tag{11}$$