Multiallelic calling model in bcftools (-m)

Petr Danecek, Richard Durbin

Version: December 5, 2013

Let f_A be an estimate of the allele frequency of the base A across all samples

$$f_A = \frac{\sum_k q_k \delta_{b_k,A}}{\sum_k q_k},$$

where $b_k \in \{A,C,G,T\}$ is the base at k-th read and q_k the corresponding quality[1] and

$$\begin{aligned} f_{A|AC} =& \frac{f_A}{f_A + f_C} \\ f_{A|ACG} =& \frac{f_A}{f_A + f_C + f_G}. \end{aligned}$$

Calculate likelihoods of all possible combinations of alleles[2] across all samples i as

$$\begin{aligned} L_A &= \prod_i L^i_A \\ L_{AC} &= \prod_i L^i_{AC} \\ L_{ACG} &= \prod_i L^i_{ACG} \\ L_{ACGT} &= \prod_i L^i_{ACGT} \end{aligned}$$

where

$$\begin{aligned} L^i_A =& PL^i(AA) = P(data|AA) \\ L^i_{AC} =& f^2_{A|AC} PL^i(AA) + f^2_{C|AC} PL^i(CC) + 2 f_{A|AC}\, f_{C|AC} PL^i(AC) \\ L^i_{ACG} =& f^2_{A|ACG} PL^i(AA) + f^2_{C|ACG} PL^i(CC) + f^2_{G|ACG} PL^i(GG) \\ & + 2(f_{A|ACG}\, f_{C|ACG} PL^i(AC) + f_{A|ACG}\, f_{G|ACG} PL^i(AG) + f_{C|ACG}\, f_{G|ACG} PL^i(CG)) \\ L^i_{a_1,\ldots,a_n} =& \sum_j f^2_{a_j|a_1,\ldots,a_n} PL^i(a_j a_j) + 2 \sum_{j<k} f_{a_j|a_1,\ldots,a_n} f_{a_k|a_1,\ldots,a_n} PL^i(a_j a_k). \end{aligned}$$

[1] See bcf_call_glfgen in bam2bcf.c for calculation of q_k.

[2] In the current implementation, at most tri-allelic sites are considered.

Select the most likely set of alleles $\{a\}$ with likelihood $L_{\{a\}}$ and the second most likely set of alleles $\{b\}$ with likelihood $L_{\{b\}}$. Accept $\{a\}$ either if the number of alleles in $\{a\}$ is smaller than in $\{b\}$ or if the significance level for a 1 degree of freedom likelihood ratio test

$$\chi^2 = 2 \log \frac{L_{\{b\}}}{L_{\{a\}}}$$

exceeds a given threshold. Assuming HWE, for i-th sample we select the genotype $\{a\} = a_1 a_2$ which maximizes the likelihood

$$P^i_{\{a\}} = (2 - \delta_{a_1,a_2}) f_{a_1|\{a\}} f_{a_2|\{a\}} PL^i(a_1 a_2).$$

The corresponding genotype quality is

$$\mathrm{GQ}^i_{\{a\}} = -10 \log \left[1 - \frac{P^i_{\{a\}}}{\sum_{\{b\}} P^i_{\{b\}}} \right]$$

and the call quality is

$$\mathrm{QUAL} = -10 \log \frac{L_{\{ref\}}}{\sum_{\{b\}} L_{\{b\}}}$$

for variant calls and

$$\mathrm{QUAL} = -10 \log \left[1 - \frac{L_{\{ref\}}}{\sum_{\{b\}} L_{\{b\}}} \right]$$

for non-variant calls, where the sum iterates over all possible genotypes $\{b\}$.