Amplicon Based Somatic Variant Calling

Edmund Lau

1 Background

We are shall call point base substitution type mutation only.

2 Per locus error estimate

Let

$$\mathcal{A} = \{A, T, G, C\}$$

denote the set of possible alleles[1] at any particular genomic locus of interest

$$l \in \{\text{genomic loci sequenced}\} \subset \mathbb{N}$$

and let

$$\begin{aligned} e : \mathcal{L} &\to [0,1] \\ l &\mapsto e_l = \mathbb{P}\left(\text{read is error} \,|\, \text{position is } l\right). \end{aligned}$$

We assume that the probability of an error read having allele $a \in \mathcal{A}$ to be uniform across $\mathcal{A}$, i.e.

$$\mathbb{P}_l(a \,|\, error) = \frac{1}{4}$$

3 Somatic variant VAF estimate

Somatic variant calling is performed at per locus level. Fix a locus $l \in \mathcal{L}$. We define

$$\begin{aligned} &\hat{a} \in \mathcal{A} && \text{(the reference allele at } l\text{)} \\ &n = (n_a)_{a\in\mathcal{A}} = (n_A, n_T, n_G, n_C) && \text{(the count of each allele, or pileup, at } l\text{)} \\ &v = (v_a)_{a\in\mathcal{A}} = (v_A, v_T, v_G, v_C) && \text{(the allele frequency at } l\text{)} \\ &V = \sum_{a\in\mathcal{A}} v_a. && \text{(the sum of the (non-error) allele frequency)} \end{aligned}$$

Now, the probability of observing allele $a \in \mathcal{A}$ at position l is given by

$$\begin{aligned} \mathbb{P}_l(a) &= \mathbb{P}_l(a \,|\, error)\mathbb{P}_l(error) + \mathbb{P}_l(a \cap \overline{error}) && (1) \\ &= \frac{1}{4}e_l + v_a && (2) \end{aligned}$$

which give the log-likelihood function in terms of the non-reference allele frequency given the empirical allele counts $n = (n_a)_{a\in\mathcal{A}}$ as

$$L(v \,|\, n) = \sum_{a\in\mathcal{A}} n_a \log\left(\frac{1}{4}e_l + v_a\right) \tag{3}$$

3.1 Maximum Likelihood using KKT optimisation

For maximum likelihood prediction, we want to find

$$v^* = \underset{v\in[0,1]^3, V\leqslant 1}{\operatorname{argmax}} L(v \,|\, n).$$

[1] The preferred ordering is (A, T, G, C)

In other words, we have the optimisation problem: maximise $L(v \mid n)$, subjected to

$$\begin{aligned} -v_a &\leqslant 0 && \forall a \in \mathcal{A} \\ V + e_l - 1 &= v_A, v_T, v_G, v_C + e_l - 1 = 0 \end{aligned}$$

the last of which is simply the constraint that (2) is a probability measure. Note that these are enough to ensure $v_a \leqslant 1$ for all $a \in \mathcal{A}$. This give rise to the Lagrangian

$$\mathcal{L}(v; \lambda, \mu) = L(v \mid n) - \mu(V + e_l - 1) + \sum_{a \in \mathcal{A}} \lambda_a v_a$$

The corresponding KKT conditions are:

$$\begin{aligned} \frac{n_a}{\frac{1}{4}e_l + v_a} &= \mu - \lambda_a && && \text{stationary, } \partial_{v_a}\mathcal{L}(v; \lambda, \mu) = 0 \\ \lambda_a v_a &= 0 && \forall a \in \mathcal{A} && \text{complimentary slackness} \\ \lambda_a &\geqslant 0 && \forall a \in \mathcal{A} && \text{dual feasibility} \\ V + e_l - 1 &= 0 && && \text{primal feasibility.} \end{aligned}$$

Stationary condition and primal feasibility jointly implies that

$$v_a = \frac{n_a}{\mu - \lambda_a} - \frac{e_l}{4} \tag{4}$$

$$\sum_{a \in \mathcal{A}} \frac{n_a}{\mu - \lambda_a} = 1 \tag{5}$$

In the case where v is in the interior of the domain, $v \in (0,1)^4$, we have for all $a \in \mathcal{A}$, $v_a > 0 \implies \lambda_a = 0$, which gives,

$$v_a = \frac{n_a}{n} - \frac{e_l}{4}$$

where $n = \sum_{a \in \mathcal{A}} n_a$ is the total allele count at position l. In particular, if $e_l = 0$, we have the solution to the uncorrected multinomial maximum likelihood predictor $p_a = n_a/n$.
We now turn to the case where v lies in one of the lower dimensional boundaries (faces, edges and corners). Let k be the number of $a \in \mathcal{A}$ such that $v_a = 0$. Observe that $\mu - \lambda_a = \frac{4n_a}{e_l}$ whenever, $v_a = 0$ and $\lambda_a = 0$ whenever $v_a \neq 0$ as dictated by the complementary slackness condition. Thus, (5), gives

$$\begin{aligned} \sum_{a \in \mathcal{A}} \frac{n_a}{\mu - \lambda_a} = \sum_{v_a = 0} \frac{e_l}{4} + \sum_{v_a \neq 0} \frac{n_a}{\mu} &= 1 \\ \implies \mu = \frac{\sum_{v_a \neq 0} n_a}{1 - \frac{k e_l}{4}}&. \end{aligned}$$

3.2 Likelihood ratio test

Under the null-hypothesis that there is no variant at position l, the probability of observing allele $a \in \mathcal{A}$ is given by

$$\mathbb{P}_l(a) = \begin{cases} 1 - \frac{3}{4}e_l & , a = \hat{a} \\ \frac{1}{4}e_l & , a \neq \hat{a} \end{cases}$$

where $\hat{a} \in \mathcal{A}$ is the reference allele at position l. Thus, the log-likelihood function is given by

$$L_0(n) = n_{\hat{a}} \log\left(1 - \frac{3}{4}e_l\right) + \log\left(\frac{1}{4}e_l\right) \sum_{a \in \mathcal{A} \setminus \hat{a}} n_a.$$

The likelihood ratio test statistics is then given by

$$G = 2 \times (L(v^* \mid n) - L_0(n))$$

and its distribution is given by χ^2 distribution with 3 degrees of freedom[2].

[2]The degree of freedom for the null model is zero while that of the alternative is 3, given by 4 parameters which sum to a given constant.