

# soma-snv: Background and Derivations

Mark Pinese

June 28, 2018

Consider the problem of identifying somatic SNVs from NGS data. Non-reference bases at a locus may be due to a subclonal SNV, machine error, poorly-sampled heterozygous sites, or mapping artefacts. The primary goal is to efficiently distinguish true subclonal SNV events from these other possibilities, and maintain good SNR even with relatively low sequencing depth. A secondary goal is to quantify the ‘callability’ of each locus, to permit normalisation of the detected somatic variant rate to the number of bases examined.

## A simple SNV detector

For simplicity, we consider each locus independently, and aim to only identify somatic variants at homozygous reference germline loci. Let the germline genotype at the locus be  $G \in \{RR, RA, AA\}$ , where  $R$  represents the reference allele and  $A$  all possible non-reference alleles, collapsed to a single symbol. Represent the observed allelic read depths by  $n_R$  and  $n_A$ , and the total depth by  $n = n_R + n_A$ ; with corresponding random variables  $N_R$ ,  $N_A$ , and  $N$ . Suppose that the sequencing process has a fixed per-base error rate  $\epsilon$ , where an error means that an incorrect base has been reported by the sequencer (ie in the case of an error, the sequencer will never report the correct base).

Express the tail probability of at least the observed number of  $A$  alleles being due to machine error at a homozygous reference locus as

$$\begin{aligned} Pr(\text{err}) &= Pr(N_A \geq n_A \mid N = n, G = RR) \\ &= \sum_{i=n_A}^n Binom(i; n, \epsilon) \end{aligned}$$

and similarly the tail probability that the counts could be due to a poorly-sampled het locus, assuming unbiased allele sampling, as

$$\begin{aligned} Pr(\text{het}) &= Pr(N_A \leq n_A \mid N = n, G = RA) \\ &= \sum_{i=0}^{n_A} Binom(i; n, \frac{1}{2} + \frac{1}{3}\epsilon) \end{aligned}$$

Given the above, we define a simple SNV caller that reports a subclonal SNV at a locus if the following is satisfied:

$$CALL \equiv (Pr(err) < \alpha_E) \wedge (Pr(het) < \alpha_H)$$

with  $\alpha_E$  an error threshold for a false positive due to noise at a reference locus, and  $\alpha_H$  an error threshold for a false positive due to poor sampling at a heterozygous locus.

Informally, this simple variant caller identifies variants which have too many  $A$  alleles to be plausibly due to machine error, yet too few to be plausibly due to a poorly-sampled heterozygous locus.

### Adaptive filtering thresholds

The performance of the simple caller described above critically depends on the values of  $\alpha_E$  and  $\alpha_H$ . Here we derive a method to determine values for these thresholds that are adaptive to local sequencing depth and user-specified parameters, and achieve optimal sensitivity for a given limit on false positive rate.

Let the rate of heterozygous loci be  $r_h$ , and the fraction of  $RR$  loci with subclonal variation be  $r_c$ . We proceed to characterise the expected sensitivity  $p_n$  and SNR  $p_r$  to detect a subclonal variant of allele frequency  $f$ .

Define performance metrics as

$$p_r \equiv \frac{Pr(SCV \mid CALL)}{Pr(\neg SCV \mid CALL)}$$

$$p_n \equiv Pr(CALL \mid SCV)$$

with  $SCV$  denoting the true presence of a subclonal variant at the locus. Expanding the above,

$$\begin{aligned} Pr(SCV \mid CALL) &= \frac{Pr(CALL \mid SCV)Pr(SCV)}{Pr(CALL)} \\ &= \frac{Pr(CALL \mid SCV)r_c}{Pr(CALL)} \\ Pr(\neg SCV \mid CALL) &= \frac{Pr(CALL \mid \neg SCV)Pr(\neg SCV)}{Pr(CALL)} \\ &= \frac{Pr(CALL \mid \neg SCV)(1 - r_c)}{Pr(CALL)} \\ \implies p_r &= \frac{Pr(CALL \mid SCV)}{Pr(CALL \mid \neg SCV)} \frac{r_c}{1 - r_c} \end{aligned}$$

The conditional call probability at a true subclonal locus can be simplified

$$\begin{aligned}
Pr(CALL \mid SCV) &= r_{RR} Pr(CALL \mid SCV, G = RR) \\
&+ r_{RA} Pr(CALL \mid SCV, G = RA) \\
&+ r_{AA} Pr(CALL \mid SCV, G = AA) \\
&= r_{RR} Pr(CALL \mid SCV, G = RR)
\end{aligned}$$

with  $r_{RA} = r_H$ ,  $r_{RR} = \frac{1}{2}(1 - r_H + \sqrt{1 - 2r_H})$ , and  $r_{AA} = \frac{1}{2}(1 - r_H - \sqrt{1 - 2r_H})$  the marginal probabilities of each genotype class. The second and third terms are zero by construction: we have defined subclonal variants to only exist at homozygous reference loci ( $G = RR$ ). Note that  $r_c$  should be chosen to take this restricted definition into account. Similarly,

$$\begin{aligned}
Pr(CALL \mid \neg SCV) &= r_{RR} Pr(CALL \mid \neg SCV, G = RR) \\
&+ r_{RA} Pr(CALL \mid \neg SCV, G = RA) \\
&+ r_{AA} Pr(CALL \mid \neg SCV, G = AA) \\
&\approx r_{RR}\alpha_E + r_{RA}\alpha_H
\end{aligned}$$

where the first term equals  $r_{RR}\alpha_E$  and the second  $r_{RA}\alpha_H$  by construction, and the third term is approximately zero (as a  $G = AA$  locus will almost surely fail the one-sided heterozygous locus exclusion filter, and due to the  $r_{AA}$  term will be rare besides).

Expanding the conditional true detection probability at a homozygous reference germline locus,

$$Pr(CALL \mid SCV, G = RR) = \sum_{n_A=0}^n \binom{n_A}{n} p_A^{n_A} (1 - p_A)^{n - n_A} [CALL]$$

with

$$p_A = \left(1 - \frac{4}{3}\epsilon\right) f + \epsilon$$

The Iverson brackets in the above can be removed by observing that the  $[CALL]$  term defines bounds on the summation of the  $n_A$  allele count distribution: let  $c_E \equiv \inf\{n_A : Pr(\text{err}) < \alpha_E\}$  and  $c_H \equiv \sup\{n_A : Pr(\text{het}) < \alpha_H\}$ , then

$$\begin{aligned}
&CALL \\
&\implies Pr(\text{err}) < \alpha_E \wedge Pr(\text{het}) < \alpha_H \\
&\implies \sum_{i=n_A}^n Binom(i; n, \epsilon) < \alpha_E \wedge \sum_{i=0}^{n_A} Binom(i; n, \frac{1}{2}) < \alpha_H \\
&\implies n_A \geq c_E \wedge n_A \leq c_H \\
&\implies c_E \leq n_A \leq c_H
\end{aligned}$$

and so

$$Pr(CALL \mid SCV, G = RR) = \sum_{n_A=c_E}^{c_H} \binom{n}{n_A} p_A^{n_A} (1 - p_A)^{n-n_A}$$

These  $c_E$  and  $c_H$  are effectively critical values of the respective tests.

Combining the above, we have

$$p_n = r_{RR} \sum_{n_A=c_E}^{c_H} \binom{n}{n_A} p_A^{n_A} (1 - p_A)^{n-n_A}$$

$$p_r = \frac{r_c}{1 - r_c r_{RR} \alpha_E + r_{RA} \alpha_H} p_n$$

### Optimisation

The goal is identify thresholds  $\alpha_E$  and  $\alpha_H$  that yield optimal sensitivity  $p_n$  for a given  $n$  and  $f$ , subject to SNR  $p_r$  being at least as good as a minimum acceptable  $g_r$ . This optimisation is performed by the following method:

1. Perform a scan along the  $c_E = c_H$  axis to identify an initial search starting point  $(c_E, c_H)$  where  $p_r \geq g_r$ .
2. Assess  $p_r$  and  $p_n$  at candidate points  $(c_E - 1, c_H)$ ,  $(c_E, c_H - 1)$ , and  $(c_E - 1, c_H - 1)$ . If at least one of the candidate points have  $p_r \geq g_r$  and has improved  $p_n$ , update the current point to this best point and repeat. Else, terminate.

As  $f$  is pre-specified and  $n$  takes on relatively few values, a lookup table containing previously-identified optima makes this simple strategy computationally efficient.

### Power and callability

The  $p_r$  values yielded by the optimisation algorithm above represent estimates of per-sample and locus sensitivity to detect subclonal variants. These  $p_r$  values can be summed across loci in a sample, possibly subset by sequence context, to yield a measure of total callability for that sample. This measure is a suitable normalisation constant to derive per-sample subclonal SNV rates.

### Input data and parameters

Given the simplicity of the error model, only high quality data should be used. In particular, read mapping quality and base quality thresholds are encouraged, and problematic loci (eg non-unique mapping) should be excluded. Preliminary work uses thresholds of  $MQ \geq 40$ ,  $BQ \geq 30$ , and excludes poor mapping or low complexity regions.

Reasonable parameters for the caller operating on high-depth sequence data ( $\geq 100\times$ ) are  $r_h = 10^{-3}$ ,  $r_c = 10^{-6}$ ,  $\epsilon = 10^{-2}$ ,  $g_r = 10$ . These must be relaxed for more typical data ( $\approx 30\times$ ).