Supplementary Methods for "Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs."

Christopher T. Saunders, Wendy Wong, Sajani Swamy,
Jennifer Becq, Lisa J. Murray and R. Keira Cheetham

March 28, 2012

1 Supplementary Methods

1.1 Strelka workflow

1.1.1 Realignment Search

The alignment search used during Strelka's realignment step is based on the assumption that a segment of each read is already aligned correctly and the alignments only need to be adjusted based on a sparse set of candidate indels. The search proceeds as follows. Each read has a starting alignment provided in the BAM input, as well as a trial set of indels. This trial set is composed of intersecting candidate indels in addition to any non-candidate 'private' indels from the starting alignment. A set of alignments is built from the starting alignment by recursively toggling indels from the trial set. Each indel is toggled from an existing parent alignment to create two child alignments where the previous alignment is preserved on the left or right side of the toggled indel. New candidate indels are added to the trial set when intersected by child alignments extending beyond the range of their parent. This search process is efficient for the common case where a read intersects only one or two candidate indels. Due to exponential complexity with the trial set size, heuristics are used to curtail the search in regions with high candidate indel density. Specifically, the search recursion depth is limited to no more than 5, and will be set lower such that the number of enumerated alignments could not exceed 5000.

1.2 Somatic variant calling model

1.2.1 Single sample SNV likelihood

The sample SNV likelihood is computed from the set of basecalls aligned to each site following the tier-specific read filtration and realignment steps described in Methods. In addition to the filtration that takes place on entire reads, an

extra filtration step is applied on the basecalls within a read, referred to as the mismatch density filter. This filter removes a basecall from consideration if more than M mismatches occur between the read and the reference within a window of 41 bases. This 41 base window is typically centered at the site in question unless restricted by the edge of the read, in which case it extends 41 bases into the read from the edge. Note that each indel counts as a single mismatch for this filter. The mismatch threshold M is set to 3 when calling at tier1 and 10 at tier2.

The SNV likelihood is computed from the remaining basecalls at the site D_x, where x refers to either the tumor or normal sample. By treating each basecall as an independent observation of the site, the likelihood can be computed as

$$P(D_x|f_x) = \prod_{b \in D_x} \sum_{a \in A} P(b|a)P(a|f_x)$$

where A is the allele set (the 4 nucleotides), $P(b|a)$ is the probability of observing basecall b given the true base a and $P(a|f_x)$ is the sampling probability of base a, a value equal to the frequency of allele a in f_x. Given basecall error probability e, the observation error term is

$$P(b|a) = \begin{cases} 1-e & \text{if } b = a \\ e/3 & \text{otherwise} \end{cases} \quad (1)$$

Strelka relies on the tumor-normal subtraction process and the explicit representation of noise allele frequencies to account for error dependencies at each site, so no further corrections are made to the independent observation model.

To represent the strand-bias states of the SNV allele frequency space, the likelihood is computed separately assuming the strand bias occurs on the forward and reverse strand. Each strand-bias component likelihood is computed with the bias strand using the allele frequency indicated in f_x and the non-bias strand fixed to a reference allele frequency of 1 (i.e. non-reference basecalls observed on the non-bias strand are penalized as basecalling error). The final strand-bias likelihood is the mean of the two strand-specific likelihoods $P_{\text{bias}}(D_x|f_x) = (P_{\text{bias}}(D_x|f_{x,\text{forward}}) + P_{\text{bias}}(D_x|f_{x,\text{reverse}}))/2$.

1.2.2 Single sample indel likelihood

As described in the realignment section in Methods, each read is processed during realignment into a set of alignments L, with probabilities computed for each alignment $l \in L$ of $P(l) = \prod_{(b,a) \in l} P(b|a)$, where b and a are the observed and expected basecalls respectively at each matched position in the alignment, and $P(b|a)$ is described in equation 1. The expected basecalls correspond to either the reference sequence or the inserted sequence recorded with each candidate insertion allele.

The indel likelihood is computed from the set of reads intersecting the indel. A read intersects the indel if it has intersecting alignments which include and exclude the indel in question (at least one of each type). The most probable

alignments from each of the indel-including and indel-excluding alignment sets are used in the reduced alignment set $T = \{i, n\}$. Note that the indel-excluding alignment may correspond to the reference sequence or another overlapping indel. While this enables overlapping indel alleles to be called, we must add a correction to filter out cases where substantial support exists for more than two alleles (described in the following section).

Each intersecting read makes an independent contribution to the indel likelihood, thus

$$P(D_x|f_x) = \prod_{d \in D_x} P(d|f_x)$$

where $P(d|f_x)$ is the single read indel likelihood. This likelihood incorporates a term to reduce the contribution of noisy read alignments to the final result by defining mapping state $M = \{m, o\}$ to indicate reads which are respectively correctly and incorrectly mapped. Incorporating this state the read likelihood is $P(d|f_x) = P(d|T, M)P(T|f_x)P(M)$. The prior for correct read mapping is $P(m) = \frac{1}{2N}$ for genome size N, and the alternate alignments are considered independent of the mapping state, so the read likelihood can be expressed as

$$P(d|f_x) = P(d|o)(1 - P(m)) + P(d|T)P(T|f_x)P(m) \quad (2)$$

where the mismapped read alignment term is $P(d|o) \approx (\frac{1}{4})^n$ for read length n. The mapping qualities provided by the read mapper are not incorporated into the indel likelihood because Strelka is designed to optionally accept local *de-novo* assembly contigs for which mapping quality is unavailable, although this functionality is not used in the present study.

For the remaining terms in equation 2, $P(d|T)$ is the alignment probability (equivalent to the $P(l)$ term described above) for each of the two alignments in the reduced alignment set T, and $P(T|f_x)$ incorporates a term to account for spurious indels as a function of homopolymer length. It is computed as $P(T|f_x) = \sum_{v \in V} P(T|v)P(v|f_x)$ where $V = \{i, n\}$ represents the true alignment path among the reduced alignment set possibilities. Here, the true alignment probability $P(v|f_x)$ is simply the frequency of the indel or reference allele in f_x, while the relationship between the observed and true indel alleles is empirically based as follows: we define $P(T|v) = 1 - e^{-f_{\text{si}}(Z,h)}$, where f_{si} is the spurious indel function dependent on indel state $Z = \{\text{insertion}, \text{deletion}\}$ and homopolymer length h. All indels besides homopolymer expansions and contractions are assigned homopolymer length 1. The spurious insertion probability is

$$f_{\text{si}}(\text{insertion}, h) = a_1 h + a_2 h^{a_3}$$

with $a_1 = 5.038 \times 10^{-7}$, $a_2 = 3.306 \times 10^{-10}$ and $a_3 = 6.998$. The spurious deletion probability is

$$f_{\text{si}}(\text{deletion}, h) = \begin{cases} b_0 & \text{if } h = 1 \\ b_1 h + b_2 h^{b_3} & \text{if } h > 1 \end{cases}$$

with $b_0 = 3.001 \times 10^{-6}$, $b_1 = 1.098 \times 10^{-5}$, $b_2 = 5.197 \times 10^{-10}$, and $b_3 = 6.993$.

1.2.3 Overlapping indel filtration

As described above, in the process of evaluating each indel we choose the highest scoring alignment which includes and excludes the indel. Usually the alignment which excludes the evaluated indel will be the reference allele, but the protocol does allow for the presence of an overlapping indel in this alternate alignment, providing an approximate method for overlapping indel calling. Many of these overlapping indel sites correspond to spurious indel hotspots, so additional filtration is used to reduce sites where the alternate indel alleles suggest more than two haplotypes. The filtration scheme enumerates an approximate measure of support for each allele S_a by summing the probability of each allele for a given read over all reads. We then require that the allele in question have an S_a value among the two highest for all overlapping alleles at the locus, and that the total S_a value of the top two alleles is at least 90% of the total for the top three. The allele probabilities for each read used to compute S_a are taken from the posterior probability over all alternate indel alleles for a given read, where this posterior is computed from the alternate alignment likelihoods for each allele $P(l)$ assuming a uniform prior probability over the set of indel alleles.

1.2.4 Normal sample diploid genotype distribution

As described in Methods, the final quality score provided by Strelka for each variant is the joint probability of a somatic variant and the most likely normal diploid genotype from the posterior genotype distribution $P(G_n|D_n)$. This distribution is computed from the same normal sample likelihoods $P(D_n|f_n)$ described for SNVs and indels in the sections above. We note that the diploid genotype states G_n are a subset of the normal sample allele frequencies f_n where frequency values are restricted to $\{0, 0.5, 1\}$, and therefore describe the diploid genotype posterior as a distribution over the normal allele frequencies where the probability of any frequency not in G_n is zero. The normal sample allele frequency posterior is

$$P_{\text{diploid}*}(f_n|D_n) \propto P(D_n|f_n)P_{\text{diploid}*}(f_n)$$

For indels, the prior allele frequency distribution used for this case $P_{\text{diploid}*}(f_n)$ is the same as the term $P_{\text{diploid}}(f_n)$ previously described in Methods. For SNVs, a variant 'polymorphic' site prior is used. As in the previous definition, $P_{\text{diploid}*}(f_n)$ is defined for the canonical diploid allele frequencies and 0 otherwise. It can be described in terms of $\alpha = f_n(a_{\text{ref}})$, the frequency of the reference allele and $\beta = f_n(a_2)/f_n(a_1)$, the allele frequency ratio of the first and second most frequent alleles, a_1 and a_2, respectively. For SNVs this is

$$P_{\text{diploid}*,\text{SNV}}(f_n) = \begin{cases} (1-\theta_{\text{SNV}})/6 & \text{if } \alpha = 0.5 \\ (1-\theta_{\text{SNV}})/12 & \text{if } \alpha = 0, \beta = 0 \\ \theta_{\text{SNV}}/3 & \text{if } \alpha = 0, \beta = 0.5 \\ (1-\theta_{\text{SNV}})/4 & \text{if } \alpha = 1 \end{cases}$$

where the heterozygosity is $\theta_{\text{SNV}} = 1 \times 10^{-3}$.