A Likelihood Approach to Somatic Mid-Size Indel Calling Under Alignment and Typing Uncertainty

Louis Dijkstra

May 23, 2014

1 Introduction

1.1 Notation

2 Methods

The methods presented in this section are based on the following assumptions:

(A1) Reads are independent with respect to their alignment uncertainties, i.e., the probability that a read originates from a certain position on the genome does not depend on the alignment probability of another read;

(A2) The read generating process does not favor particular chromosomes, i.e., a read is a priori equally likely to stem from any chromosome present in the sample;

(A3) The alignment quality of a read is not (heavily) influenced by the presence/absence of the variant allele.

We will throughout the text note explicitly when these assumptions come into play. Note that assumption *(A2)* entails that a variant-allele-affected chromosome is equally likely to bring forth a read as a chromosome that does not harbor the allele.

2.1 Variant allele frequency (VAF) estimation

Let $R = \{R_1, R_2, \ldots, R_m\}$ be the set of reads. Every read $R_i \in R$ is associated with two events:

1. A_i^R – the read R_i stems from the locus/region of interest, and
2. V_i^R – the read R_i stems from a variant-allele-affected chromosome,

where the superscript R refers to the read's sample (this notation becomes useful when we start to reason about several read samples simultaneously, see Section XXX). The probability $\mathbb{P}\left\{A_i^R\right\}$ is commonly provided by the aligner. The conditional probability $\mathbb{P}\left\{V_i^R \mid A_i^R\right\}$, i.e., the chance that i-th read comes from a variant-allele-affected chromosome given that the read stems from the locus of interest, can often be estimated, see Section XXX. Throughout the rest of the paper, we will use a_i^R and v_i^R as a shorthand

notation for $\mathbb{P}\left\{A_i^R\right\}$ and $\mathbb{P}\left\{V_i^R \mid A_i^R\right\}$. For the rest of this section, these quantities are assumed to be known for every read.

Let us in addition define for every subset $S \subset \{1, \ldots, m\}$ the event A_S^R that only the reads R_i with $i \in S$ originate from the locus of interest, while all other reads in R do not. Since reads are independent w.r.t. to their alignment probabilities *(A1)*, its probability is given by

$$\mathbb{P}\left\{A_S^R\right\} = \prod_{i \in S} \mathbb{P}\left\{A_i^R\right\} \prod_{j \notin S} \left(1 - \mathbb{P}\left\{A_j^R\right\}\right) = \prod_{i \in S} a_i^R \prod_{j \notin S} \left(1 - a_j^R\right). \tag{1}$$

Let $\phi \in \Phi \subset [0, 1]$ be the variant allele frequency[1] (VAF), i.e., the fraction of haplotypes that carry the variant allele. In this section, we set out to estimate this quantity using maximum likelihood given the read data in R and the alignment and typing uncertainties as expressed by the probabilities $\left\{(a_i^R, v_i^R)\right\}_{i=1}^m$.

The likelihood fucntion $\mathcal{L} : \Phi \to [0, 1]$ returns the probability of observing data R in the case of a VAF equal to ϕ, i.e., $\mathcal{L}(\phi) = \mathbb{P}_\phi\{R\}$. (We use $\mathbb{P}_\phi\{\cdot\}$ to explicitly denote when the probability measure depends on the parameter ϕ). Since only the reads in R that truely stem from the locus of interest can provide information about ϕ, the likelihood function can be written as

$$\mathcal{L}(\phi) = \sum_{S \subset \{1,\ldots,m\}} \mathbb{P}_\phi\left\{A_S^R\right\} \mathbb{P}_\phi\left\{R \mid A_S^R\right\}, \tag{2}$$

where $\mathbb{P}_\phi\left\{R \mid A_S^R\right\}$ is the probability to observe R given that only the reads R_i with $i \in S$ stem from the locus and the other reads do not. By assumption *(A3)* the probability $\mathbb{P}_\phi\left\{A_S^R\right\} = \mathbb{P}\left\{A_S^R\right\}$ (reads stemming from a variant-allele-affected chromosome are equally likely to be aligned correctly as other reads) and is, therefore, given by eq. (1). Since reads are assumed to be independent *(A1)*, one can write

$$\mathbb{P}_\phi\left\{R \mid A_S^R\right\} = \prod_{i \in S} \mathbb{P}_\phi\left\{R_i \mid A_i^R\right\} \prod_{j \notin S} \mathbb{P}_\phi\left\{R_j \mid \overline{A_j^R}\right\} \tag{3}$$

where $\mathbb{P}_\phi\left\{R_i \mid A_i^R\right\}$ and $\mathbb{P}_\phi\left\{R_i \mid \overline{A_i^R} i\right\}$ are the probabilities of observing read R_i given that it either stems (A_i^R) or does not stem ($\overline{A_i^R}$) from the locus of interest. Let us first derive $\mathbb{P}_\phi\left\{R_i \mid A_i^R\right\}$. Under assumption *(A2)* read R_i stems from a variant-allele-affected chromosome with probability equal to the VAF ϕ (and, similarly, does not descend from a chromosome carrying the variant allele with probability $1 - \phi$). Due to typing uncertainties, one commonly only knows whether the read stems from a variant-allele-affected chromosome up to a certain level v_i^R. The probability to observe a read R_i given that it stems from the locus of interest is, therefore,

$$\begin{aligned} \mathbb{P}_\phi\left\{R_i \mid A_i^R\right\} &= \phi \cdot \mathbb{P}\left\{V_i^R \mid A_i^R\right\} + (1 - \phi)\left(1 - \mathbb{P}\left\{V_i^R \mid A_i^R\right\}\right) \\ &= \phi \cdot v_i^R + (1 - \phi)\left(1 - v_i^R\right). \end{aligned} \tag{4}$$

[1]The possible values of the VAF ϕ in Φ can often be restricted to a finite set, e.g., for diploid organisms $\Phi = \{0, 1/2, 1\}$, since either none, only one or two chromosomes can harbor the variant allele. In some cases, it can be rather unclear how to pick Φ; for tumor samples the number of distinct haplotypes are often unknown. Working with the unit interval $\Phi = [0, 1]$ can then be a pragmatic solution. We will return to the choice of Φ in more extent later in the paper.

The probability $\mathbb{P}_\phi\left\{R_j \mid \overline{A_j^R}\right\} = \mathbb{P}\left\{R_j \mid \overline{A_j^R}\right\}$ does not depend on ϕ since the variant allele frequency only tells us something about the reads stemming from the locus of interest. Since the read must stem from somewhere on the genome, the probability of observing this read given that it does not cover our region is equal to 1.

Combining these results yields an expression for the likelihood of $\phi \in \Phi$ given the read data R and the alignment and typing uncertainties $\left\{(a_i^R, v_i^R)\right\}_{i=1}^m$:

$$\mathcal{L}(\phi) = \sum_{S \subset \{1,\ldots,m\}} \prod_{i \in S} a_i^R \left[\phi \cdot v_i^R + (1-\phi)\left(1 - v_i^R\right)\right] \prod_{j \notin S} \left(1 - a_j^R\right). \tag{5}$$

Computing this likelihood would on first glance requires summing over all possible subsets of reads, which entails exponential runtime in the number of reads. Fortunately, this expression can be rewritten as

$$\begin{aligned} \mathcal{L}(\phi) &= \prod_{i=1}^m \left\{a_i^R \left[\phi \cdot v_i^R + (1-\phi)\left(1 - v_i^R\right)\right] + \left(1 - a_i^R\right)\right\} \\ &= \prod_{i=1}^m \left[\phi \cdot a_i^R \left(2v_i^R - 1\right) + 1 - a_i^R v_i^R\right] \end{aligned} \tag{6}$$

which is linear in the number of reads.

2.1.1 Genotyping: the di- and polyploid case

$\phi \in \Phi_c = \{0, 1/c, 2/c, \ldots, 1\}$.

$$\widehat{\phi}_c = \max\{\mathcal{L}_{\mathrm{c}}(0), \mathcal{L}_{\mathrm{c}}(1/c), \mathcal{L}_{\mathrm{c}}(2/c), \ldots, \mathcal{L}_{\mathrm{diploid}}(1)\}. \tag{7}$$

$\phi \in \Phi_{\mathrm{diploid}} = \{0, 1/2, 1\}$.

$$\mathcal{L}_{\mathrm{diploid}}(\phi) = \begin{cases} \prod_{i=1}^m \left(1 - a_i^R v_i^R\right) & \text{if } \phi = 0, \\ \prod_{i=1}^m \left(1 - \frac{a_i^R}{2}\right) & \text{if } \phi = 1/2, \\ \prod_{i=1}^m \left(1 - a_i^R v_i^R - a_i^R\right) & \text{if } \phi = 1. \end{cases} \tag{8}$$

$$\widehat{\phi}_{\mathrm{diploid}} = \max\{\mathcal{L}_{\mathrm{diploid}}(0), \mathcal{L}_{\mathrm{diploid}}(1/2), \mathcal{L}_{\mathrm{diploid}}(1)\}. \tag{9}$$

2.1.2 The continuous case

The previous approach of enumerating all possible values of ϕ and computing and comparing their respective likelihoods becomes for obvious reasons infeasible when we allow ϕ to vary over the unit interval.

The loglikelihood function ℓ is easily found to be

$$\ell(\phi) = \log \mathcal{L}(\phi) \sum_{i=1}^m \log\left\{\phi \cdot a_i^R \left(2v_i^R - 1\right) + 1 - a_i^R v_i^R\right\}. \tag{10}$$

We find that its second order derivative is

$$\frac{\partial^2 \ell}{\partial \phi^2} = -\sum_{i=1}^{m} \left[\frac{a_i^R \left(2v_i^R - 1\right)}{\phi \cdot a_i^R \left(2v_i^R - 1\right) + 1 - a_i^R v_i^R} \right]^2 < 0 \tag{11}$$

for all $\phi \in [0, 1]$ given that there is at least one read that potentially stems from the locus of interest and provides information on the presence/absence of the variant allele, i.e., $\exists i \in \{1, \ldots, m\}$ for which $a_i^R > 0$ and $v_i^R \neq 1/2$. Therefore, the loglikelihood function is, under this weak condition, a strictly concave function, and therefore possesses one unique global maximum on the unit interval.

2.2 Likelihood construction for somatic mutation calling

Suppose we are presented with two matched data sets of reads, $C = \{C_1, \ldots, C_m\}$ and $D = \{D_1, \ldots, D_n\}$, where C is the control sample of (presumed) healthy, diploid cells and D is the disease sample: a set of reads that stem from a mixture of tumor and healthy cells. Let $\alpha \in [0, 1]$ denote the fraction of chromosomes in this disease-control mixture D that are from healthy cells, commonly referred to as the *level of impurity*. For the moment we assume α to be known.

In order to assess whether a somatic mutation occured, we need to determine the frequency with which the variant allele occurs in both healthy and tumor cells. We will denote these quantities with, respectively, ϕ_h and ϕ_t (where the 'h' reffers to healthy and 't' to tumor). In the case that the variant is absent in healthy cells, i.e., $\phi_h = 0$, and present in tumor cells, i.e., $\phi_t > 0$, we speak of a somatic mutation.

Since healthy cells are diploid, the values ϕ_h can take are contained in $\Phi_h = \{0, 1/2, 1\}$, since either zero, one or two chromosomes can harbor the allele. The frequencies with which the variant can occur among tumor cells are harder to establish. Ideally, one would know the number of distinct haplotypes present in the tumor. Unfortunately, estimates on this number are hard to obtain. We, therefore, choose to vary ϕ_t over the unit interval, i.e., $\Phi_t = [0, 1]$, thereby avoiding the need to specify this quantity. Although it might be superfluous to treat a discrete quantity such the VAF as a continuous one, we feel that assuming continuity will not heavily undermine the performance of the method presented here.

More formally, we are interested in estimating the parameter $\vartheta = (\phi_h, \phi_t) \in \Theta$, where

$$\Theta = \Phi_h \times \Phi_t = \{(\phi_h, \phi_t) : \phi_h \in \{0, 1/2, 1\}, \phi_t \in [0, 1]\} \tag{12}$$

is our parameter space. The likelihood function $\mathcal{L}_{\text{SM}} : \Theta \to [0, 1]$ is the probability of observing the case and control data sets given the parameter ϑ, i.e., $\mathcal{L}_{\text{SM}}(\phi_h, \phi_t) = \mathbb{P}_{\phi_h, \phi_t}\{C, D\}$. Due to the independence of the control and case sample, this probability can be written as

$$\mathcal{L}_{\text{SM}}(\phi_h, \phi_t) = \mathbb{P}_{\phi_h}\{C\}\mathbb{P}_{\phi_h, \phi_t}\{D\} \tag{13}$$

where we used the fact that the healthy sample C is not influenced by the tumor VAF ϕ_t. Note that due to impurity, quantified by α, the disease sample D does provide information about the VAF of healthy cells, ϕ_h. The first term $\mathbb{P}_{\phi_h}\{C\}$ is identical to the likelihood of ϕ_h in the diploid case, see eq. (XXX). Determining $\mathbb{P}_{\phi_h, \phi_t}\{D\}$ is slightly more involved, but similar to the derivation of the likelihood function for the VAF in Section XXX. Again,

only the reads in D that actually stem from the locus of interest carry information about ϑ, i.e.,

$$\mathbb{P}_{\phi_h,\phi_t}\{D\} = \sum_{S\in\{1,\dots,n\}} \mathbb{P}\left\{A_S^D\right\} \mathbb{P}_{\phi_h,\phi_t}\left\{D \mid A_S^R\right\} \tag{14}$$

where we used assumption *(A3)* and $\mathbb{P}\left\{A_S^D\right\}$ is given in eq. (1). Due to the independence of reads *(A1)*, we can write

$$\mathbb{P}_{\phi_h,\phi_t}\left\{D \mid A_S^R\right\} = \prod_{i\in S} \mathbb{P}_{\phi_h,\phi_t}\left\{D_i \mid A_i^D\right\} \prod_{j\notin S} \mathbb{P}_{\phi_h,\phi_t}\left\{D_j \mid \overline{A_j^D}\right\} \tag{15}$$

where $\mathbb{P}_{\phi_h,\phi_t}\left\{D_j \mid \overline{A_j^D}\right\}$ is again equal to 1. Under assumption *(A2)*, the read R_i stems with probability α from a healthy cell and with probability $1-\alpha$ from a tumor cell. Therefore,

$$\mathbb{P}_{\phi_h,\phi_t}\left\{D_i \mid A_i^D\right\} = \alpha\mathbb{P}_{\phi_h}\left\{D_i \mid A_i^D\right\} + (1-\alpha)\mathbb{P}_{\phi_t}\left\{D_i \mid A_i^D\right\}. \tag{16}$$

The probabilities $\mathbb{P}_{\phi_h}\left\{D_i \mid A_i^D\right\}$ and $\mathbb{P}_{\phi_t}\left\{D_i \mid A_i^D\right\}$ are given by eq. (4). The probability $\mathbb{P}_{\phi_h,\phi_t}\{D\}$ is, therefore,

$$\begin{aligned}\mathbb{P}_{\phi_h,\phi_t}\{D\} &= \sum_{S\in\{1,\dots,n\}} \prod_{i\in S} a_i^D\left[\alpha\mathbb{P}_{\phi_h}\left\{D_i \mid A_i^D\right\} + (1-\alpha)\mathbb{P}_{\phi_t}\left\{D_i \mid A_i^D\right\}\right] \prod_{j\notin S}\left(1-a_i^D\right) \\ &= \prod_{i=1}^{n}\left\{a_i^D\left[\alpha\mathbb{P}_{\phi_h}\left\{D_i \mid A_i^D\right\} + (1-\alpha)\mathbb{P}_{\phi_t}\left\{D_i \mid A_i^D\right\}\right] + \left(1-a_i^D\right)\right\}.\end{aligned} \tag{17}$$

Combining these results yields the following expression for the likelihood $\mathcal{L}_{\text{SM}}(\phi_h,\phi_t)$:

$$\begin{aligned}\mathcal{L}_{\text{SM}}(\phi_h,\phi_t) &= \mathbb{P}_{\phi_h}\{C\}\mathbb{P}_{\phi_h,\phi_t}\{D\} \\ &= \prod_{i=1}^{m}\left[a_i^C\phi_h\left(2v_i^C-1\right)+1-a_i^Cv_i^C\right]\times \\ &\qquad \prod_{i=1}^{n}\left[a_i^D\left(\alpha\phi_h+(1-\alpha)\phi_t\right)\left(2v_i^D-1\right)+1-a_i^Dv_i^D\right].\end{aligned} \tag{18}$$

$$\mathcal{L}_{\text{SM}}(\phi_h,\phi_t) = \begin{cases} \prod_{i=1}^{m}\left(1-a_i^Cv_i^C\right)\times\prod_{i=1}^{n}\left[a_i^D\phi_t(1-\alpha)\left(2v_i^2-1\right)+1-a_i^Dv_i^D\right] & \text{if } \phi_h=0 \\ \prod_{i=1}^{m}\left(1-\frac{a_i^C}{2}\right)\times\prod_{i=1}^{n}\left[a_i^D\left(\phi_t(1-\alpha)+\frac{\alpha}{2}\right)\left(2v_i^D-1\right)+1-a_i^Dv_i^D\right] & \text{if } \phi_h=1/2 \\ \prod_{i=1}^{m}\left(1-a_i^Cv_i^C-a_i^C\right)\times\prod_{i=1}^{n}\left[a_i^D\left(\phi_t(1-\alpha)+\alpha\right)\left(2v_i^D-1\right)+1-a_i^Dv_i^D\right] & \text{if } \phi_h=1 \end{cases} \tag{19}$$

$$\frac{\partial^2\ell(\vartheta)}{\partial\phi_t^2} = -\sum_{i=1}^{n}\left[\frac{a_i^D(1-\alpha)\left(2v_i^D-1\right)}{a_i^D\left(\phi_t(1-\alpha)+\alpha\phi_h\right)\left(2v_i^D-1\right)+1-a_i^Dv_i^D}\right]^2 \leq 0$$

Strictly concave iff $\alpha<1$ and $\exists i\in\{1,\dots,n\}$ for which $a_i^D>0$ and $v_i^D\neq 1/2$.

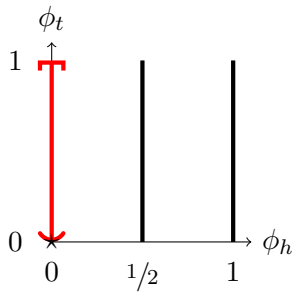

Figure 1: The parameter space Θ. The red interval $(0, 1]$ at $\phi_h = 0$ is Θ_{SM} and denotes the cases where a somatic mutation occured. The parameters related to the situation where no mutation occured are depicted with black, including the origin $(0, 0)$ which is for clarity shown here with a $\star$.

2.3 Testing for somatic mutations: A Likelihood ratio test for non-nested models

$$\Theta_{\text{SM}} = \{(\phi_h, \phi_t) : \phi_h = 0, \phi_t \in (0, 1]\}. \tag{20}$$

$\Theta_0 = \Theta \setminus \Theta_{\text{SM}}$.

$$H_0 : \vartheta \in \Theta_0 \qquad \text{versus} \qquad H_a : \vartheta \in \Theta_{\text{SM}}$$