Enhancing sensitivity and controlling false discovery rate in somatic indel discovery

Johannes Köster[1,2,3*,†], Louis J. Dijkstra[4,3*,†],
Tobias Marschall[5,6,3*,‡], Alexander Schönhuth[3,‡]

[1] Algorithms for reproducible bioinformatics, Genome Informatics,
Institute of Human Genetics, University Hospital Essen, University of Duisburg Essen, Germany
[2] Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
[3] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[4] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany
[5] Max Planck Institute for Informatics, Saarbrücken, Germany
[6] University Saarbrücken, Germany
[†] Joint first authorship
[‡] Joint last authorship
[*] Previous affiliation

alexander.schoenhuth@cwi.nl

August 21, 2019

Abstract

As witnessed by various population-scale cancer genome sequencing projects, accurate discovery of somatic variants has become of central importance in modern cancer research. However, count statistics on somatic insertions and deletions (indels) discovered so far point out that large amounts of discoveries must have been missed. The reason is that the combination of uncertainties relating to, for example, gap and alignment ambiguities, twilight zone indels, cancer heterogeneity, sample purity, sampling and strand bias are hard to accurately quantify. Here, a unifying statistical model is provided whose dependency structures enable to accurately quantify all inherent uncertainties in short time. As major consequence, false discovery rate (FDR) in somatic indel discovery can now be controlled at utmost accuracy. As demonstrated on simulated and real data, this enables to dramatically increase the amount of true discoveries while safely suppressing the FDR. Specifically supported by workflow design, our approach can be integrated as a post-processing step in large-scale projects.

The software is publicly available at `https://varlociraptor.github.io` and can be easily installed via Bioconda[1] [Grüning et al., 2018].

[1] `https://bioconda.github.io`

1 Introduction

Cancer is a genetic disorder in the first place; somatic mutations turn originally healthy cells into a heterogeneous mix of aberrantly evolving cell clones [Burrell et al., 2013]. Global consortia have launched population-scale sequencing projects concerned with the discovery and annotation of somatic variants in cancer genomes [The International Cancer Genome Consortium, 2010, Weinstein et al., 2013]. Potential benefits of the systematic analysis of somatic mutations include improved diagnosis, staging and therapy protocol selection in the clinic.

The fraction of somatic variants discovered, however, has left room for improvement across the whole range of possible variant types [Alioto et al., 2015]. Thereby, somatic insertions and deletions (indels)[2] have proven to pose particular challenges when belonging to certain classes or length ranges [Hause et al., 2016, Maruvka et al., 2017]. Indels of length approximately 30 - 250 bp, termed ”the next-generation sequencing (NGS) twilight zone of indels” in other contexts [Mandoiu and Zelikovsky, 2016, Marschall et al., 2013, Trappe et al., 2014], have resisted their discovery in particular also in somatic variant calling: while the COSMIC database[3] counts 1,879,044 indels of length 1-30 bp, it only counts 17 793 indels of length 31-60 bp, 3758 indels of length 61-100 bp and 2483 indels of length 101-250 bp. The drop by two orders of magnitude from 1-30 to 30-60 bp, not followed by any such drops in further length bins, has so far not been supported by any reasonable biological interpretation. Our benchmark experiments demonstrate that this size range still corresponds to a blind spot in somatic mutation discovery. The most likely explanation is that the majority of somatic indels in that size range have remained undiscovered so far.

Therefore, the application of more sensitive[4] somatic indel calling strategies most likely will induce striking changes in the spectrum of somatic indels so far detected. As indicated in earlier work [Maruvka et al., 2017], this has the potential to deepen our understanding of the origin and effects of somatic indels, beyond just balancing count statistics.

In this paper, we suggest such a sensitive strategy, and prove that with it we can make substantial progress in terms of eliminating the somatic indel discovery blind spot. To understand the issues that are characteristic of this blind spot, consider that somatic variant discovery (unlike germline variant discovery) is a two-step procedure: in a first step, one discovers putative variants in both the cancer and healthy (or control) genome of the individual analyzed. In the second step—which is unique to somatic variant discovery—one runs a differential analysis that classifies putative variants into somatic, because they only appear in the cancer genome, germline and healthy somatic, because variants appear in the control genome (where healthy somatic appear at subclonal levels), or just noise.

Already the first step (which also applies for generic germline variant discovery) is affected by major issues, where for example gap wander and annihilation (see [Lunter et al., 2008]) are well-known and notorious examples when determining gapped alignments in general. Issues become further aggravated when dealing with indels of 30-250 bp due to particularities of NGS read alignment and indel discovery tools. So, for determining twilight zone indels, one needs to make particular methodical efforts already when dealing with generic settings [Mandoiu and Zelikovsky, 2016, Marschall et al., 2012, 2013, Trappe et al., 2014].

In somatic indel discovery, we deal with an additional layer of issues due to cancer heterogeneity. The variant allele frequency (VAF), here the fraction of genome copies in the (tumor or control) sample affected by the variant, is either 0.0, 0.5, or 1.0 for germline variants, reflecting absence, hetero- or homozygosity. In contrast, allele frequencies of somatic variants vary across the whole range from 0.0 to 1.0, depending on the clonal structure of the tumor sample and its impurity (the ratio of healthy genome copies in the

[2] Here, we refer to insertions and deletions of all possible sizes, ranging from 1 to thousands of base pairs.

[3] `https://cancer.sanger.ac.uk/cosmic`; Release v88, data retrieved on 20th of March, 2019.

[4] Here and in the following, *sensitivity or recall* denotes the ratio of true indels discovered over the overall amount of true indels, whereas *precision* denotes the ratio of true indels discovered over the overall amount of indels discovered. *False discovery rate (FDR)* is the ratio of mistaken discoveries over the overall amount of indels discovered (so precision = 1 - FDR).

tumor sample). Usually, there is no prior information about the clonal structure available at the time of variant calling. Low-frequency variants (i.e. having a VAF close to zero) yield particularly weak, statistically uncertain signals. Of course, there are limits to somatic variant discovery, relative to VAF and sequencing depth. The methodical challenge is to not miss any discoveries that a statistically sound approach is able to reveal. The purpose of this paper is to do this: we would like to fathom the corresponding limits theoretically and to push them considerably in practice.

Certainly, there are still also improvements for the first step conceivable; however, the second (differential analysis) step has never been treated before with statistical rigor. So, the second step may have left room for improvements in particular. We recall that dealing with the various statistical uncertainties due to (as above-mentioned) cancer heterogeneity, gap placement, strand bias, etc., is the major challenge in the second step. For successful operation, one needs to accurately quantify all relevant uncertainties in the first place. As a consequence, one has a sound statistical account on whether putative variants are somatic, germline, healthy somatic, or just errors.

Key to success in increasing the number of true somatic variant discoveries is to establish statistically sound *false discovery rate (FDR)* control, as the canonical procedure to limit the number of false predictions when increasing the output. In an ideal setting, the user specifies the maximal ratio of false discoveries s/he is willing to deal with. Subsequently, a maximal set of discoveries is reported that ensures the FDR specified by the user. It is a well-known insight that accurate FDR control only works if uncertainties have been accurately quantified: otherwise inaccuracies propagate towards FDR control, implying insufficient control, too little discoveries, or both. This explains why working with accurately quantified uncertainties is imperative when seeking to safely increase the amount of true discoveries.

The desired accuracy in uncertainty quantification, however, comes at a cost: if data is uncertain, one deals with an exponential amount of possibly true data scenarios. This prevents naive approaches to work at the desired level of accuracy without serious runtime issues, which constitutes a common computational bottleneck in uncertainty quantification. In our setting, we will be dealing with at least 3^n possible scenarios for one putative indel where n corresponds to the sum of read coverages in the cancer and control genome at a particular locus. Nowadays, $n \leq 60$ is not uncommon. This number of arithmetic operations is prohibitive when processing up to hundreds of thousands putative indel loci.

In the model presented in this paper, we overcome this bottleneck by presenting a Bayesian latent variable model whose conditional dependencies point out how to compute all of the relevant probabilities in time *linear, and not exponential* in the coverage of a putative indel locus. Thanks to its computational efficiency, the model captures and accurately quantifies all uncertainties involved in somatic indel discovery in a comprehensive manner.

In summary, we are able to compute all probabilities required for enforcing reliable FDR control at both the necessary speed and accuracy. Reliable, sound and accurate FDR control, in turn, gets us in position to substantially increase the recall in somatic indel discovery, without having to deal with losses or even improving in terms of precision. As was to be expected, improvements in the twilight zone turn out to be most dramatic: in comparison with state-of-the-art tools we double or even triple recall, while preserving (often better than just) operable precision.

2 Results

We have designed *Varlociraptor*, as a method to implement the improvements in the differential analysis step outlined in the Introduction. To the best of our knowledge, Varlociraptor is the first method that allows for accurate and statistically sound false discovery rate control in the discovery of somatic indels. As a consequence, the application of Varlociraptor leads to substantial increases in recall in somatic indel discovery. Varlociraptor doubles or even triples the number of true discoveries in comparison with state-of-the-art tools, while often also further improving on precision, or, at any rate, not incurring any kind of loss in precision. Varlociraptor also accurately estimates the variant allele frequency (VAF) for all somatic indels.

In the following, we provide a high-level description of the Varlociraptor workflow. We provide a brief explanation of how one can quantify all relevant uncertainties in linear runtime, as the key methodical breakthrough and the fundamental building block that underlies all that follows. We briefly illustrate how the Bayesian latent variable model that enables rapid quantification of uncertainties further immediately gives rise to the computation of all probabilities that are crucial in somatic variant calling. We further briefly address how Varlociraptor estimates variant allele frequencies (VAF's). Finally, we explain how accurate FDR can be established. For details, we refer to the Methods section.

Subsequently, we analyze Varlociraptor's performance in comparison with current state-of-the-art tools on simulated and real data. As pointed out above, we show that Varlociraptor indeed achieves (sometimes drastic) increases in recall, often accompanied by further increases in precision. We notice that the probabilities used for classifying putative variant calls allow for a clear distinction between true and false positives, which is of considerable value in classification practice. We then demonstrate that Varlociraptor indeed reliably controls FDR, thereby also providing the theoretical explanation for why Varlociraptor achieves superior performance rates in terms of recall and precision. Varlociraptor further accurately estimates all VAF's. Turning our attention to real data, we conclude that Varlociraptor achieves superior concordance for variants of VAF at least 20%. For variants of VAF of less than 20%, Varlociraptor is the only tool that discovers considerable amounts of variants. The low coverage of reads supporting such calls delivers stringent statistical explanations for why concordance cannot be reached at rates that apply for calls above 20%. To corroborate that the majority of Varlociraptor's calls are correct—just as we experienced on simulated data—we demonstrate that Varlociraptor's count statistics agree with the theoretical expectation under neutral evolution.

2.1 Workflow

We first discuss how Varlociraptor embeds in a workflow for somatic variant calling and highlight the central difference to classical approaches.

The classical workflow for calling somatic variants (Figure 1a) starts with aligned reads from tumor and corresponding healthy sample of the same patient in BAM[5] or CRAM[6] format. First, variants are discovered and a differential analysis is performed to call variants as somatic or germline. Candidate variants are reported in VCF or BCF format[7]. In the following we will refer to VCF as a placeholder for VCF/BCF, and BAM as a placeholder for BAM/CRAM. Second, the candidate variants are filtered, usually by applying thresholds for various scores (e.g. variant quality, strand bias, coverage, minimum mapping quality, minimum number of supporting reads in healthy sample), in order to obtain final variant calls. If not just relying on some suggested defaults, finding those thresholds is often a tedious, study specific effort.

With Varlociraptor, we provide a new approach for calling and filtering, thereby separating variant discovery from calling (Figure 1b). The input for Varlociraptor are candidate variants from an external discovery step. Here, any variant calling tool can be applied. [8] There are plenty of approaches that thoroughly deal with the discovery step. In particular, as it is already common practice within state of the art somatic vari-

[5] https://samtools.github.io/hts-specs/SAMv1.pdf

[6] https://samtools.github.io/hts-specs/CRAMv3.pdf

[7] https://samtools.github.io/hts-specs/VCFv4.3.pdf

[8] At the time of writing, the Varlociraptor implementation supports SNVs, insertions and deletions. However, the model presented here is agnostic of the variant type and we are actively working on adding support for all other types of variants in Varlociraptor. Therefore, while we are writing about indels in the following, keep in mind that the presented model can straightforwardly be applied to other variant types. Similarly, Varlociraptor currently supports single-end or paired-end short reads, while the model itself is agnostic of the sequencing protocol and technology. The implementation will be extended in the future.

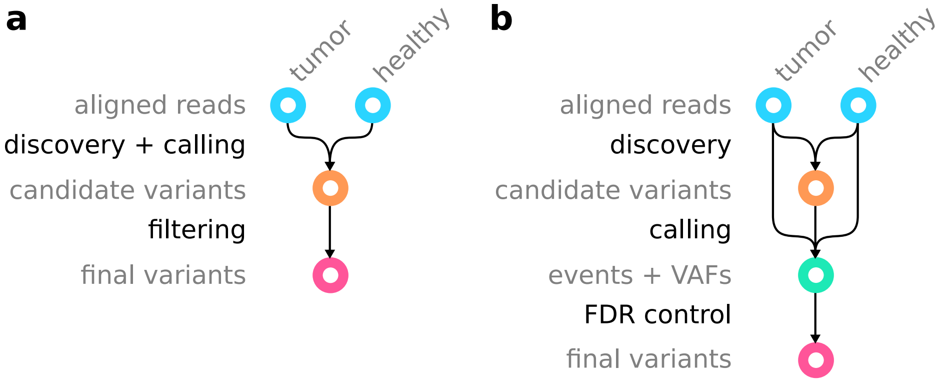


Figure 1: Difference between (a) the classic somatic variant calling workflow and (b) the Varlociraptor approach.

ant calling pipelines (e.g. the Sarek pipeline[9]), it is possible to combine the candidate variants of different callers in order to obtain maximum sensitivity across all variant types and length ranges. However, instead of having to perform ad-hoc merging of finalized calls (e.g. Sarek[9] performs majority voting), Varlociraptor provides a unified mechanism for assessing all candidate variants. During calling, Varlociraptor classifies variants into somatic tumor, somatic healthy, germline or absent variants, while providing posterior probabilities for each event (see section 2.2.2) along with maximum a posteriori estimates of the variant allele frequency (VAF) (see section 2.2.3), reported in BCF format. Finally, using the posterior probabilities, Varlociraptor can filter variants by simply controlling for a desired false discovery rate (FDR, see section 2.2.4), instead of requiring the adjustment of various thresholds. This becomes possible because Varlociraptor, as the first approach, integrates all known sources of uncertainty into a single, unified model.

In this work, we will evaluate Varlociraptor's performance in direct comparison with the (usually ad-hoc) differential analysis routines provided by other tools. We will use the indels that are output by the competitor as per its first (discovery) step as input for Varlociraptor. This way we ensure that all tools receive input they are supposed to deal with, by their design, and therefore ensure maximum fairness.

2.2 Foundation of the Approach

2.2.1 Efficient Computation of the Fundamental Likelihood Function

Let us fix a particular variant locus, as given by an entry in the VCF file that lists all candidate variants from Figure 1b. By θ_h and θ_c we denote the true, but unknown allele frequency of that (putative) variant among the healthy (θ_h) and the cancerous (θ_c) genome copies. While for *germline* variants $\theta \in \{0, {}^1\!/\!{}_2, 1\}$ reflecting absence (artifacts or noise), hetero- and homozygosity of the variant, $\theta \in [0, 1]$ for somatic variants. We model that *somatic healthy* variants usually show at subclonal rates by allowing only $\theta \in (0, {}^1\!/\!{}_2)$, i.e. the exclusive interval between 0 and ${}^1\!/\!{}_2$. A variant evaluates as *somatic tumor* if and only if $\theta_c > 0$, while $\theta_h = 0$. Since we are most interested in these variants, one of the central goals is to conclude that $\theta_c > 0, \theta_h = 0$ for a particular putative variant with sufficiently large probability.

By $\boldsymbol{Z}^h = (Z^h_1, ..., Z^h_k)$ and $\boldsymbol{Z}^t = (Z^t_1, ..., Z^t_l)$, we denote the read data being associated with the variant locus in the healthy (h) and the tumor (t) sample[10]. Each of the $Z^h_i, Z^t_j, i = 1, ..., k, j = 1, ..., l$ represents one (paired-end) read that became aligned across or nearby the given variant locus. This further means that k

[9] https://github.com/SciLifeLab/Sarek

[10] Note that we distinguish between the tumor sample, which is a mixture of healthy and cancer cells, and the cancer cells themselves. When referring to the latter, we use the subscript c, for the former, we use the subscript t.

and l correspond to the sample-specific read coverages at that locus. For selecting reads via alignments, we use BWA-Mem [Li and Homer, 2010] in the following, although the choice of particular aligner is optional, as long as the aligner outputs a MAPQ value [Li et al., 2008b], which quantifies the certainty by which the sequenced fragment (represented by the read pair) stems from the locus under consideration.

Let further $\beta \in \{0, \frac{1}{2}, 1\}$ denote the *strand bias* affecting the particular variant locus. Thereby $\beta = 0$ and $\beta = 1$ denote that evidence about the putative variant occurs only in the reverse ($\beta = 0$) or in the forward ($\beta = 1$) strand. Both cases are indicators of sequencing or mapping artifacts. Therefore, no strand bias, i.e. $\beta = {}^1\!/_2$ will subsequently be used to select for non-artifact variants.

We present a *Bayesian latent variable model* that enables to *efficiently compute*

$$L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t), \tag{1}$$

the likelihood of allele frequencies θ_h, θ_c and strand bias β given read data $\boldsymbol{Z}^h, \boldsymbol{Z}^t$ (see section 5.3). Straightforward approaches to computing (1) via *fully Bayesian inverse uncertainty quantification* [Liu et al., 2009], which is the canonical and approved way to computing (1) fail due to requiring exponential runtime. The conditional dependency structure of the statistical model we raise (see Methods, section 5.3), points out a way to compute (1) in runtime linear in $k + l$, as summarized by the following theorem.

Theorem 2.1. *$L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ can be computed in $O(k + l)$ arithmetic operations.*

Note that this is the best one can hope for; the insight is crucial in somatic variant calling practice, beyond establishing also a theoretical novelty, because Theorem 2.1 establishes that $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ can be evaluated in short runtime at any $(\theta_h, \theta_c, \beta)$. This in turn renders integration over $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ computationally feasible, which facilitates solving the following three essential problems.

2.2.2 Classification

Statistically sound classification in somatic variant calling requires, when given $\boldsymbol{Z}^h, \boldsymbol{Z}^t$, to compute the posterior probabilities for the following four cases, which refer to different combinations of θ_c, θ_h (see also Figure 8a in Methods).

- somatic tumor (st); $\theta_c > 0, \theta_h = 0$: the variant is somatic in the tumor, and does not appear in the healthy genome,
- germline (ge); $\theta_h \in \{{}^1\!/_2, 1\}$: a germline variant, where $\theta_h = {}^1\!/_2$ reflects a heterozygous and $\theta_h = 1$ a homozygous variant,
- somatic healthy (sh); $\theta_h \in (0, {}^1\!/_2)$: a variant that is somatic but appears in the healthy genome, reflected by subclonal, non-germline variant allele frequencies, or
- absent (ab); $\beta \in \{0, 1\}$ or $\theta_c = 0, \theta_h = 0$: the variant reflects (strand bias) artifacts or noise

Note in particular that cases (st), (ge), (sh) imply that there is no strand bias, i.e. $\beta = {}^1\!/_2$. Given the respective read data $\boldsymbol{Z}^h, \boldsymbol{Z}^t$ from the healthy and the tumor genome, the corresponding posterior probabilities compute as

$$P\left(\text{st} \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t\right) = \frac{1}{P(\boldsymbol{Z}^h, \boldsymbol{Z}^t)} \int_0^1 h(0, \theta_c, {}^1\!/\!_2) L(0, \theta_c, {}^1\!/\!_2 \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t) d\theta_c, \tag{2}$$

$$P\left(\text{ge} \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t\right) = \frac{1}{P(\boldsymbol{Z}^h, \boldsymbol{Z}^t)} \int_0^1 \sum_{\theta_h \in \{1/2, 1\}} h(\theta_h, \theta_c, {}^1\!/\!_2) L(\theta_h, \theta_c, {}^1\!/\!_2 \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t) d\theta_c, \tag{3}$$

$$P\left(\text{sh} \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t\right) = \frac{1}{P(\boldsymbol{Z}^h, \boldsymbol{Z}^t)} \int_0^{1/2} \int_0^1 h(\theta_h, \theta_c, {}^1\!/\!_2) L(\theta_h, \theta_c, {}^1\!/\!_2 \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t) d\theta_c d\theta_h, \tag{4}$$

$$P\left(\text{ab} \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t\right) = 1 - \sum_{\{st, ge, sh\}} P\left(\cdot \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t\right). \tag{5}$$

$h(\theta_h, \theta_c, \beta)$ is the prior distribution of the three readouts $\theta_h, \theta_c, \beta$, which can be used to integrate prior knowledge about clonal structure or zygosity rates, if available. We consider the choice of h as an open question, that is most important for sparse data (i.e. very low coverage) It can be guided by the work of Williams et al. [2016], for example. For the evaluation conducted here, we use a uniform h. $P(\boldsymbol{Z}^h, \boldsymbol{Z}^t)$ is the marginal probability of the data, which acts as a normalization factor.

The integrals lack an analytic formula, but, supported by the efficient computation of (1), as established by theorem 2.1, can be approximated numerically using quadrature.

2.2.3 Estimating allele frequencies for somatic tumor variants

Upon having determined that a variant is somatic tumor, implying $\theta_c > 0, \theta_h = 0$ and $\beta = \frac{1}{2}$, we would like to determine maximum a posteriori estimates for θ_c, given the fragment data $\boldsymbol{Z}^h, \boldsymbol{Z}^t$. When using a uniform prior this is the same as computing the maximum likelihood estimate

$$\widehat{\theta_c} \equiv \operatorname*{arg\,max}_{\theta_c \in [0,1]} L(0, \theta_c, {}^1\!/\!_2 \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t). \tag{6}$$

The likelihood function (1) is a higher-order polynomial in θ_h and θ_c for given β, as follows from the computations in section 5.3.4, which makes it infeasible to derive its maximum analytically. We can nevertheless prove the following theorem.

Theorem 2.2. *For fixed θ_h and β, the logarithm of the likelihood function $\theta_c \to L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ is concave on the unit interval $\mathcal{U} = [0, 1]$. Hence the likelihood function attains a unique global maximum $\widehat{\theta_c}$ on $[0, 1]$.*

This allows to report the corresponding global maximum $\widehat{\theta_c}$ for $\theta_h = 0$ and $\beta = \frac{1}{2}$, as a reasonable estimate for the VAF of a somatic cancer variant. See Theorem A.1 in Appendix A for a detailed technical exposition, including a proof and a list of extra conditions for this theorem necessary to hold, all of which apply in practice. Since the logarithm of the likelihood function is strictly concave, its maximum can be easily determined numerically.

2.2.4 False Discovery Rate (FDR) Control

Once the event probabilities defined in section 2.2.2 are available, FDR can be controlled in a statistically sound way. In an ideal setting, the user specifies an FDR threshold γ, upon which a maximal amount of variants is output such that the ratio of false discoveries among the discoveries overall does not exceed γ. Only if the underlying model is statistically sound, maximal increases in terms of true discoveries among the output (sensitivity or recall) can be expected. Note that ad-hoc style FDR control procedures such as 'call merging' (raising only discoveries simultaneously supported by several variant callers), while indeed controlling FDR, usually incur serious losses in terms of discoveries, and tend to lead to discovery blind spots, thereby (while not being the main issue) contributing to the lack of 'somatic twilight zone indels' so far discovered. Moreover, it becomes next to impossible to fine-tune an analysis to a particular FDR acceptable in the specific context. First tries on FDR control procedures for variant calling have been reported before [DePristo et al., 2011, Marschall et al., 2012]. However, in this work, we present the first fully Bayesian approach, and also the first for the calling of somatic variants.

For a given set of putative somatic tumor variants C, each of which is annotated with $p_i, i = 1, \dots, |C|$, the posterior probability (2) that variant i is indeed somatic tumor, we can calculate the expected FDR [Mueller et al., 2004] as

$$FDR_C = \frac{1}{|C|} \sum_{i=1}^{|C|} 1 - p_i$$

In order to both control FDR at γ and raise maximal output, one searches for the largest set of variants $C^* \subseteq C$ such that $FDR_{C^*} \leq \gamma$. One can efficiently implement this by sorting variants by $1 - p_i$ in ascending order, and summing up $1 - p_i$ in that order until this sum divided by the number of summands collected, has reached the threshold γ.

2.3 Data analysis reproducibility

The evaluation performed in this paper is available as a reproducible Snakemake [Köster and Rahmann, 2012] workflow archive[11]. In addition we provide a *Snakemake report* that allows to interactively explore all Figures shown in this article in the context of the workflow, the parameters, and the code used to generate them (see supplementary files).

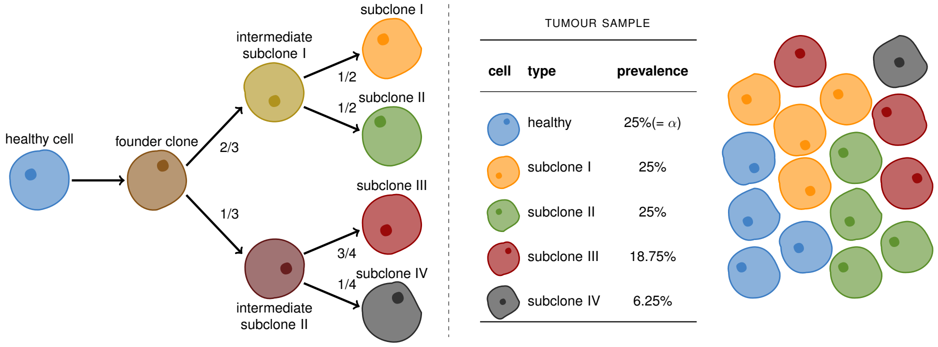
2.4 Data

A central challenge when evaluating somatic indel calling is to sufficiently cover all relevant length ranges. Furthermore, the occurrence of already discovered true twilight zone indels is biased towards easily accessible, hence lesser uncertain regions of the human genome. However, somatic variant databases contain only few twilight zone indels (see section 1). Hence, we chose a dual approach based on designing a simulated dataset with the desired properties for assessing precision and recall, complemented by a concordance analysis on real data.

Simulated Data. We used a real genome (Venter's genome [Levy et al., 2007]), previously approved for NGS benchmarking purposes [Marschall et al., 2012, 2013] as the control genome.

Our goal was to simulate a cancer genome that qualifies for statistically sound benchmarking. We therefore sampled 300 000 somatic point mutations, 150 000 insertions and 150 000 deletions, of which 279 509, 139 491 and 139 532 in the autosomes, following the clonal structure described by Figure 2, so as to reflect a scenario that is realistic in terms of tumor evolution. To account for realistic proportions in terms of length, indels follow the length distribution of Venter's germline indels (which roughly follows a power-law distribution). Allele frequencies range from 0.042 to 0.667 for the somatic indels. The majority of indels is of relatively low frequency and they were randomly placed across the genome. Both low frequencies (making indels hard to distinguish from noise) and random placement (rendering the calling particularly difficult in repetetive regions), lead to a particularly challenging benchmarking dataset.

[11] `https://doi.org/10.5281/zenodo.3361700`



cell	type	prevalence
●	healthy	25%(= α)
●	subclone I	25%
●	subclone II	25%
●	subclone III	18.75%
●	subclone IV	6.25%

(a) Simulated tumour evolution

(b) Simulated tumour sample

Figure 2: Simulated Cancer Clones: **(a)** The evolution of the cancer clones. **(b)** The simulated tumour sample. Each cell is assumed to be diploid. Cells of the same type share the same genetic code. The relative prevalences of the various cell types are shown in the table. The level of impurity (α) is 25%.

Reads were sampled using the Assemblathon read simulator SimSeq [Earl et al., 2011], at 30x and 40x for control and cancer genome, respectively. Subsequently, reads were aligned using BWA-MEM [Li and Durbin, 2009].

Real Data. To demonstrate the applicability in practice, we also evaluated real cancer-control genome pairs. Since, as of today, there are no real datasets with known ground truth available (where the lack of real twilight zone indel discoveries is of course an important factor) we opt for a concordance analysis, as a procedure that has been approved earlier. Namely, we analyzed the concordance of reported variants with respect to four replicates of cancer-control genome pairs, all of which have been sampled from the same tumor cell line (Melanoma cell line COLO829), albeit in different institutions [Craig et al., 2016][12]. If discoveries referring to the four replicates agree to a sensible degree (considering that differences due to batch effects and independent progression are conceivable), one can conclude that performance is also of high quality on real datasets.

2.5 Tools

For generating lists of candidate indels in form of VCF files (see Figure 1), we chose Delly 0.7.7 [Rausch et al., 2012], Lancet 1.0.0 [Narzisi], Strelka 2.8.4 [Saunders et al., 2012], Manta 1.3.0 [Chen et al., 2016], and BPI 1.5[13] without applying additional filtering. All these callers provide their own "ad-hoc" method of annotating somatic variants from tumor/normal sample pairs, which we applied for comparison. We provided BWA-MEM alignments with marked duplicates (via Picard-Tools[14]) as input for all tools. When subsequently running Varlociraptor, we used the output VCF files of the tools in combination with the BAM files that were the basis for generating the candidate variants.

2.6 Experiments

We first consider the simulated dataset (see section 2.4). Here, the true somatic variants are known from the simulation procedure. To classify predicted variants of the different callers into true and false positives, we matched them against the known truth using the `vcf-match` subcommand of RBT[15] with parameters

[12] `https://ega-archive.org/datasets/EGAD00001002142`

[13] `https://github.com/hartwigmedical/hmftools/tree/master/break-point-inspector`

[14] `https://broadinstitute.github.io/picard`

[15] `https://github.com/rust-bio/rust-bio-tools`

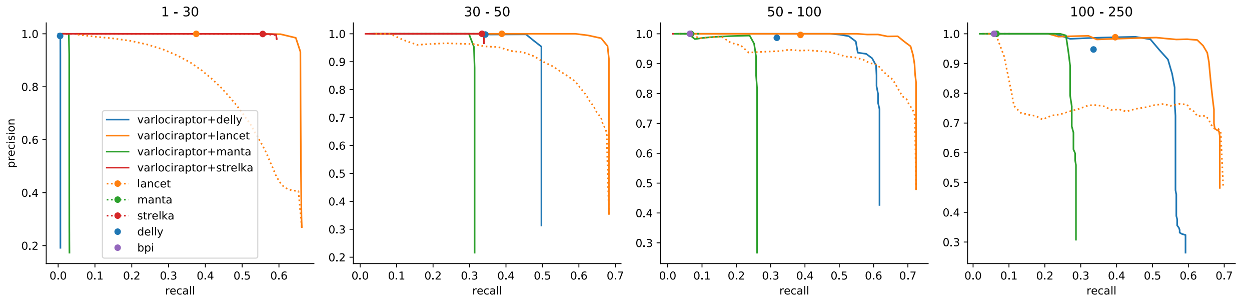


Figure 3: Recall and precision for calling somatic deletions. Results are grouped by deletion length, denoted as interval at the top of the plot. For our approach (Varlociraptor+*) curves are plotted by scanning over the posterior probability for having a somatic tumor variant. For other callers that provide a score to scan over (p-value for Lancet) we plot a dotted line. Ad-hoc results are shown as single dots. Results are shown if the prediction of the caller did provide at least 10 calls. The sharp curves for our approach reflect the favorable property of having a strong separation between the probabilities of true and false positives, see Figure 4.

`--max-dist 50 --max-len-diff 50`. This means that we consider a predicted somatic variant to be a true positive if it's position and length are within 50 bases of the true variant (which reflects approved evaluation practice, see [Wittler et al., 2015] for further reasoning). Thereby the position for a deletion is defined as its center point, i.e., $\lfloor (e-s)/2 \rfloor$ with e being the end position and s being the start position of the deletion. In the following we show the results for deletions. If results for insertions essentially deviate from the deletion results, we mention it in the corresponding section of the text. All results for insertions can be found in the supplement.

Varlociraptor achieves substantial increases in Recall without loosing Precision. In the following, *Recall* is defined to be the ratio of true variants that became predicted, while *Precision* is the ratio of correct predictions among the predictions overall. Figures 3 and S1 show Recall and Precision for all tools considered on the simulated data. Thereby, we juxtapose the tools' Recall and Precision when run in standalone modus (dots or dotted lines) with Recall and Precision when postprocessing the respective output sheets of the tools with Varlociraptor (lines). The lines referring to Varlociraptor result from varying the posterior probability threshold: the greater the threshold the smaller the Recall. While the reduction in Recall is merely a consequence of reducing the output, a simultaneous increase in precision only shows if posterior probabilities make sense, as (first) essential evidence of the quality of the approach. Note that the only caller that offers to vary an output-specific threshold is Lancet (dotted orange line in Figures 3 and S1), by scoring variants with p-values[16].

The first, fundamental observation is that Precision indeed increases on increasing the posterior probability cutoff, across all size ranges and inputs. This points out that Varlociraptor's posterior probabilities for a variant to be somatic indeed make sense.

When further comparing Varlociraptor's (continuous) lines with the dots or dotted lines referring to the standalone modus of the other tools, it becomes immediately evident that Varlociraptor improves on the tools' results: in all cases Varlociraptor's lines are upper right of the tools' dots or dotted lines. This means that Varlociraptor achieves better combinations of Recall and Precision in all cases. The most striking observation however is that Varlociraptor achieves substantial increases in recall in comparison to the standalone tools: in particular in the twilight zone (30-250 bp), Varlociraptor is able to double (Lancet, Delly) or even

[16] Lancet's p-values are only one component of the ad-hoc filtering procedure performed by the tool, which relies on multiple scores. This explains while filtering based on the p-values alone (dotted orange line) yields suboptimal performance compared to the ad-hoc calls provided by Lancet (orange dot).

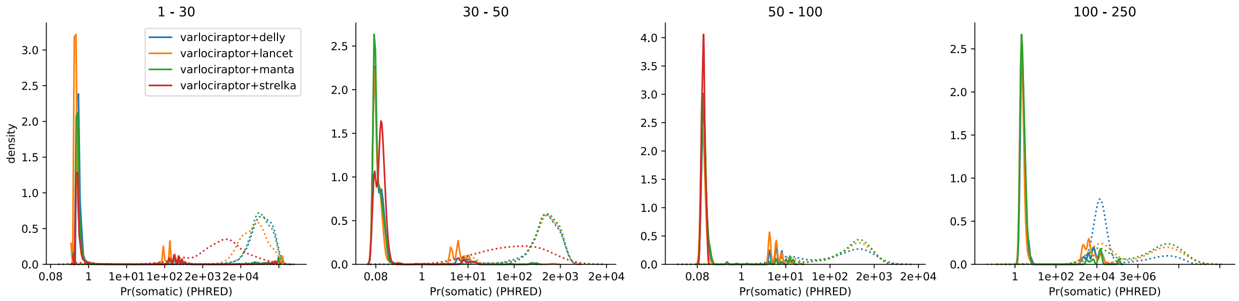


Figure 4: Posterior probability distributions for somatic deletions. Results are grouped by deletion length, denoted as interval at the top of the plot. The x-axis indicates the (PHRED-scaled) probability, and the y-axis indicates relative amounts of calls with this probability. The distributions of posteriors for true positive calls are shown as solid lines, the distributions of posteriors for false positive calls are shown as dotted lines.

more than double (Manta) the Recall, while never incurring notable losses in Precision, but usually rather improving on Precision.

Posterior probabilities allow for a clear distinction between true and false positives. The precision-recall curves of Varlociraptor show a sharp turning point at their upper right. This indicates that Varlociraptor's posterior probabilities allow for a clear and fine-grained distinction between true and false positives. Note that the vertical parts of these lines therefore also indicate the maximum recall one can achieve, the limit of which corresponds to the list of variant calls provided by the caller. In other words, Varlociraptor is able to identify all, or nearly all of the true variants that are provided as candidates. To further solidify this observation, we investigated the posterior probability distributions of Varlociraptor. Figures 4 (deletions) and S4 (insertions) show that Varlociraptor's probabilities are indeed clearly separating true from false positives.

Varlociraptor reliably controls false discovery rate (FDR). Varlociraptor summarizes the uncertainty about a putative variant in terms of a single, and, as we pointed out so far, reliable quantity: an estimate of the posterior probability for the putative variant to be somatic in the tumor. It remains to determine which of the variants to output. This is a ubiquitous issue in sequencing based variant calling: optimally, one outputs maximum amounts of variants while ensuring that the mistaken predictions among the output variants do not exceed a preferably user-defined ratio. In other words, an optimal scenario is to employ a statistically sound routine that establishes *false discovery rate (FDR) control*.

Note that establishing statistically sound FDR control in variant discovery, to the best of our knowledge, amounts to a novelty: all of the state-of-the-art methods that we benchmarked against either cannot control FDR, or establish it through ad-hoc methods, such that theoretical guarantees cannot be provided.

For Varlociraptor, it is rather straightforward to establish FDR control: well-known theory [Mueller et al., 2004] points out how to achieve maximally large output at a desired level of FDR control, when working with Bayesian type posterior probabilities. See Figure 5 (deletions) and Figure S2 (insertions) for the evaluation of how Varlociraptor's FDR control performs. In general, the closer to the diagonal, the better FDR control can be established by the user.

For deletions (Figure 5) Varlociraptor controls FDR, in the sense that across all deletion length ranges the curve is on or below the diagonal, and that deviations from the diagonal are small. The latter means that Varlociraptor tends to be slightly conservative in some combinations of length range and FDR threshold provided. One reason contributing to this is that, induced by too coarse MAPQ values or base quality scores, the resolution of posterior probabilities may be limited.

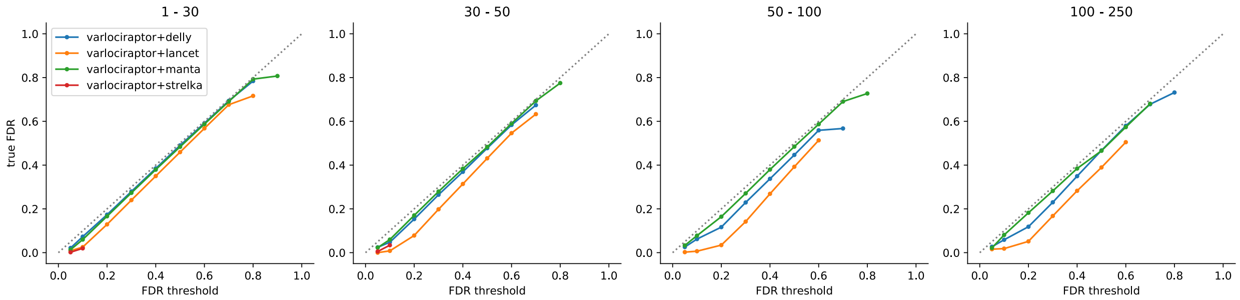


Figure 5: FDR control for somatic deletions. Results are grouped by deletion length, denoted as interval at the top of the plot. The axes denote the desired FDR, provided by the user as input (x-axis), and the true achieved FDR (y-axis). A perfect FDR control would keep the curve exactly on the dashed diagonal. Below the diagonal, the control is conservative. Above the diagonal, the FDR would be underestimated. Importantly, points below the diagonal mean that the true FDR is smaller than the threshold provided, which means that FDR control is still established; in this sense, points below the diagonal are preferable over points above the diagonal.

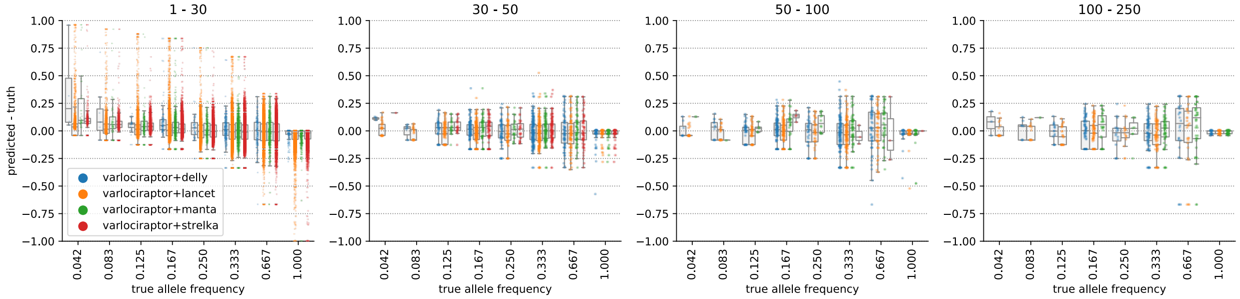


Figure 6: Allele frequency estimation for somatic deletions. Results are grouped by deletion length, denoted as interval at the top of the plot. The horizontal axis shows the true allele frequency, the vertical axis shows the error between predicted allele frequency and truth.

As for insertions (see Figure S2), which, as we recall, are generally more challenging, Varlociraptor equally achieves high-quality FDR control, across all length ranges and FDR thresholds. There is one caveat: for insertions of length 30-100 provided by Lancet [Narzisi], Varlociraptor's FDR is slightly greater than the threshold specified by the user (and, although deviations are tiny, in this sense does not control FDR). An explanation for this to happen is the modus operandi of Lancet: Lancet bases insertion calls on microassemblies that are computed from all reads mapping to the variant locus. This approach is reasonable for large insertions, which do not fit into single reads, because their length either exceeds the read length, or are too long to show in single reads at full length. However, microassembly may also lead to false positives in repetitive areas where both alignments and assemblies can lead to ambiguities; note that Lancet only reaches a precision of 90% for larger insertions. Varlociraptor, at this point in time, cannot quantify all uncertainties emerging from microassemblies—we consider it highly interesting future work to also quantify uncertainties that are associated with microassemblies (see section 3).

Varlociraptor accurately estimates variant allele frequency (VAF). Figures 6 (deletions) and S3 (insertions) show the difference between the true VAF's and the ones predicted by Varlociraptor, henceforth referred to as prediction error. The first observation is that the prediction error is approximately centered around zero in all cases, which is the desired scenario. We further investigated the effect of sequencing depth on the accuracy of the VAF estimates. Figures S6 (deletions) and S7 (insertions) show how the prediction error varies relative to sequencing depth (number of non-ambiguously mapped fragments overlapping the variant locus), denoted by n and true VAF, denoted by θ^*. For each combination of n and θ^*, we determine an expected baseline error, modeling that one samples n fragments each of which stems from a variant-affected genome copy with probability θ^*. In other words, the expected baseline error is governed by a Binomial distribution $B(n, \theta^*)$. Accordingly, we determine

$$\frac{1}{n}\sqrt{\frac{1}{n}\theta^*(1-\theta^*)},$$

that is the standard deviation of $B(n, \theta^*)$, divided (normalized) by the depth, as the expected baseline error. This establishes the theoretical optimum of a VAF estimation procedure. In other words, no sound estimator can achieve smaller error in prediction. We see that, on average, our prediction is close to this theoretical optimum. Further, the accuracy of VAF estimates increases on increasing sequencing depth, which is the desired, logical behaviour. In summary, these results point out that the estimates are sound, and even close to what one can optimally achieve in theory. A possible explanation for the remaining small deviations from the theoretical optimum are reads that stem from different loci, and because of ambiguity in placement get aligned to the variant locus. If the read mapper is unable to reflect such ambiguity in the MAPQ score, for example because the true locus of origin is due to variation not properly represented in the reference sequence, our model is as well not able to properly reflect this in the allele frequency estimation. We expect such problems to mitigate with longer reads and more accurate (even non-linear) reference genomes in the future.

Varlociraptor achieves superior concordance above VAF of 20%. We applied all callers (see section 2.5) on the four replicates of cancer-control genome pairs described above (section 2.4), using their default parameters. In all analyses, because of the lack of prior knowledge available, Varlociraptor assumed a purity level α of 100%. We then performed a concordance analysis in the following way.

For each caller we collected calls for each of the four replicates, both when run in standalone fashion and when postprocessing calls with Varlociraptor. For each of the calling strategies, we then computed matchings across the four replicates as described at the beginning of section 2.6. For each calling strategy, we then constructed a graph where each node represents one variant call in one replicate and edges indicate that two calls (from different replicates) are matched. We then consider the connected components of this graph: any non-trivial connected component (that is any connected component consisting of more than one node) counts as *concordant call*. We then determine *concordance* as the ratio of concordant calls over all connected components

In Figures 7 and S5, we display for each possible VAF, the concordance of all calls with an at least as high VAF. In other words, for each VAF threshold $t \in [0, 1]$ we display the concordance all calls with a VAF $\geq t$. To allow for a fair comparison, we use Varlociraptor's VAF estimates for all calling strategies. It becomes immediately evident that Varlociraptor achieves superior concordance for VAFs of 20% and higher.

Varlociraptor's variant counts of VAF below 20% agree with with the theoretical expectation under neutral evolution. When inspecting all calls with a minimum VAF below 20% (i.e. $t < 0.2$, see above), Varlociraptor's concordance drops below the rates achieved by callers run in standalone fashion. This does not mean, however, that Varlociraptor's performance is worse. Note first that Strelka and Lancet, which appear to have superior concordance, raise only very little (and, as we will point out below, according to evolutionary models too little) discoveries. This can be seen in the right panels of Figures 7 and S5. In contrast, Varlociraptor raises substantially more discoveries at these frequencies. Second, for low frequency

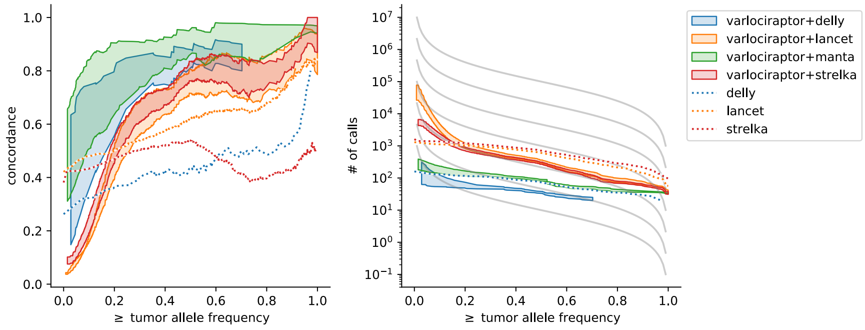


Figure 7: Concordance of somatic deletions on real data. For Varlociraptor, the interval between all calls with a posterior probability of at least 0.9 and at least 0.99 is shown as shaded area. Left: Concordance vs. minimum allele frequency. Right: Number of calls vs. minimum allele frequency. The different grey lines depict the theoretical expectation at different effective mutation rates according to Williams et al. [Williams et al., 2016] (see text).

variants, data may not reach the necessary degree of certainty (which leads to a call by Varlociraptor) in sufficiently many of the four samples, and, since the four replicates were raised in different laboratories, low frequency variants unique to samples can be expected [Craig et al., 2016].

To explore this further and assess whether the larger numbers of low frequency calls provided by Varlociraptor are potentially correct, we compared the somatic variant count distributions with the theoretical expectation under neutral evolution (see gray lines in the right panels of Figures 7 and S5). For this, we employ the tumor evolution model by Williams et al. [Williams et al., 2016], and calculate the expected number of somatic variants of at least allele frequency f as

$$M(f) = \frac{\mu}{\lambda}\left(\frac{1}{f} - \frac{1}{f_{\max}}\right).$$

Here, μ is the somatic mutation rate, λ is the growth rate, and $f_{\max}$ is the maximum clonal allele frequency. Because there are no reliable estimates for $f_{\max}$ available, we set $f_{\max} = 1.0$. We then plot the expected counts at various values of $\frac{\mu}{\lambda}$ (the effective mutation rate). It can be seen that the counts of low frequency variants provided by Varlociraptor (shaded red and orange areas) are closer to the theoretical expectation than those of Lancet and Strelka (dotted red and orange lines). This is an indication that, by evolutionary principles, it is likely that many low frequency calls that were not recognized as concordant, are still correct.

3 Discussion

We have presented a statistical framework for the calling of somatic insertions and deletions from matched tumor-normal genome samples whose application has yielded substantial increases in terms of true discoveries, while safely limiting the amount of false discoveries. The framework is implemented in an easy-to-use open source software, called Varlociraptor[17]. In comparison with the state of the art, we have demonstrated to double, or even triple the amount of true discoveries while not increasing the false discovery rate, or suppressing it even further.

We have chosen to rely on external tools for the discovery of variants and focus on providing a sound and rigorous statistical treatment of the differential analysis (or the classification) of variants into the events

[17] https://varlociraptor.github.io

relevant for somatic variant calling. An immediate benefit of this strategy is that Varlociraptor can be conveniently integrated in large-scale projects, as a post-processing step in production-quality somatic variant calling pipelines. This particularly applies when projects are run by large consortia that manage various, often heterogeneous combinations of variant callers: as per its design, Varlociraptor offers the first approach that is able to analyze sets of variants raised by different callers, all of which come with their own strengths, weaknesses, and blind spots, in a *statistically unifying* way. Because Varlociraptor preserves and combines the individual strengths of the callers while eliminating any of their particular weaknesses, it is able to raise call sheets of superior quality. In summary, the application of Varlociraptor has the potential to lead to substantial increases in true somatic indel discoveries—possibly even overwhelming in certain size ranges—in large-scale matched tumor-normal genome sequencing projects.

The technical challenge has been to overcome a well-known and notorious computational bottleneck that relates to the quantification of the uncertainties affecting the differential analysis. Thereby, not only ambiguities in terms of gap placement and aligning reads in general, which had been dealt with in the literature abundantly before, but also effects such as cancer heterogeneity (implying uncertain variant allele frequencies), purity of tumor samples, bias in terms of sampling indel-affected fragments, and strand bias, are major factors. We have presented, to the best of our knowledge, the first model that allows to capture all relevant effects and quantify all inherent uncertainties computationally efficiently. In particular, we want to stress that the presented model is the first approach that takes strand bias as a major source for systematic artifacts into a statistically comprehensive account, which is not possible if strand bias is quantified through independently raised auxiliary scores. An important aspect is the fact that the presented model can be considered as a “white box”, in the sense that all parameters have a direct biological interpretation. Hence, it supports the investigation of variant calls at different levels of detail, depending on the research question. First, one applies global filtering via FDR control. The remaining calls can optionally be investigated more closely, by looking at the estimated allele frequencies and their posterior distribution, the estimated sampling bias (see section 5.3.2 in Methods), the likelihoods and strand support of each fragment with associated mapping qualities, and finally, the read alignments themselves (if necessary). This becomes particularly important in the era of personalized medicine, where variant calls for individual patients might lead to therapy decisions, requiring upmost certainty and transparency in the decision process.

Because the statistical model comprehensively addresses all relevant effects and related uncertainties, filtering the posterior probabilities derived from the model gives immediate rise to accurate statistically sound (fully Bayesian) FDR control, for the first time in the variant calling field.

The advantage of such a straightforward and statistically interpretable filtering procedure becomes apparent when comparing it with the methodology of prior state-of-the-art approaches: so far, they have been relying on a variety of (often independently raised) scores, where combinations of (often manually) fine tuned default thresholds are supposed to ensure that predictions contain reasonably little amounts of false discoveries. Modifying these combinations of scores in order to change the default FDR, provided by the developers through the default settings, is tedious, difficult, and error-prone, if not entirely impossible.

The analysis of somatic variants often serves the purpose to yield further insight into the clonal structure of a tumor which includes to assess the allele frequencies of somatic variants appropriately. Varlociraptor, again as per its design, does not only assess the probability of individual putative somatic variants to be true discoveries, but also equips them with an estimate for their allele frequency. We have demonstrated that the corresponding VAF estimates are of almost optimal accuracy.

Still, as always, there is room for improvements. First, Varlociraptor so far only deals with insertions, deletions, and single nucleotide variants (which are not analyzed in this work). Note however the framework presented is generic in terms of its dependency structures. Therefore, efficient computation of the central likelihood function is also warranted for other variant classes; the only thing required is to adapt the computation of the typing uncertainty (see section 5.3.1 in Methods).

Second, we have been focusing on second-generation paired-end reads in this treatment, motivated by the fact that the vast majority of reads sequenced to date belongs to this class. However, our model is entirely agnostic to any particular choice of sequencing platform and can be used without any further adaptations: the only basic requirement is that the sequencing/mapping protocol in use yields (sufficiently reliable) MAPQ values. Any particular sequencing/mapping specific issues will be taken care of by the re-alignment step that Varlociraptor routinely makes use of for accurately quantifying alignment related uncertainties. When re-parameterizing the pair HMM that underlies the re-alignment step (which reflects a straightforward adaptation) to emerging long read technologies like Nanopore[18] or SMRT sequencing[19], Varlociraptor will be particularly apt for dealing with insertions and deletions within repeat regions.

Third, all quantities relating to uncertainties or spelled out probabilities that Varlociraptor (comprehensively) provides can be used further in a whole range of downstream analyses. Intriguing examples of such potential applications are the probabilistic assessment of larger somatic gains and losses, or the (partial) phasing of tumor subclones, which we will explore in future work.

Finally, it is important to note that the computational insights and the model presented in this work is, although motivated by, not at all limited to somatic variant calling. In fact, it turns out that it can be generalized towards arbitrary variant calling scenarios that require a differential analysis, where distinguishing between variants affecting the primary tumor and variants showing in metastases or relapse tumors, or distinguishing variants recurringly showing in different individual tumors from variants that do not, are immediate, relevant examples. The latest release of Varlociraptor already provides a *variant calling grammar*, as an interface for defining such scenarios[20] and thereby a foundation for a unifying theory of variant calling, enabling us to explore entirely new fields of applications.

4 Acknowledgements

We thank David Lähnemann for fruitful discussions and help with benchmarking and performance optimization. Johannes Köster was supported by NWO Veni grant 016.Veni.173.176.

5 Methods

5.1 Notation

We denote *observable variables* by Latin capital letters (e.g. Z). *Realizations* of these variables are denoted by small Latin letters (e.g. z). *Hidden/latent variables* are denoted by small Greek symbols. Vectors are denoted by boldface letters (e.g. $\boldsymbol{Z} = (Z_1, ..., Z_k)$ or $\boldsymbol{z} = (z_1, ..., z_k)$. We use super-/subscripts h and t for the healthy and the tumor sample and c to only refer to cancer cells within the tumor sample.

Let us fix a particular variant locus; we then denote the relevant read data in the healthy and the tumor sample by $\boldsymbol{Z}^h = (Z_1^h, ..., Z_k^h)$ and $\boldsymbol{Z}^t = (Z_1^t, ..., Z_l^t)$, where each of the $Z_i^h, Z_j^t, i = 1, ..., k, j = 1, ..., l$ represents one (paired-end) read that (or parts of which) became aligned across or nearby the fixed variant locus. For selecting reads via alignments, we use BWA-Mem [Li and Homer, 2010] in the following, although the choice of particular aligner is free, as long as the aligner outputs a MAPQ value, which quantifies the certainty by which the reads stem from the locus under consideration.

By *variant allele frequency (VAF)*, we refer to the fraction of genome copies in the sample affected by the variant. We denote this (unknown) frequency in the healthy and the tumor sample by θ_h and θ_t, respectively. Since healthy cells are diploid, θ_h is either 0, $1/2$, or 1 corresponding to absence, heterozygosity and homozygosity, when dealing with a *germline variant*. It is common that healthy cells, beyond germline

[18] https://nanoporetech.com

[19] https://www.pacb.com/smrt-science/smrt-sequencing

[20] https://varlociraptor.github.io/docs/calling#generic-variant-calling

variation, also exhibit somatic mutations. Somatic variants that affect healthy cells usually occur at subclonal rates, i.e. are generally not characteristic in terms of giving rise to subpopulations among the healthy cells. It is therefore reasonable to assume that somatic variants affect only less than half of the cells, reflected by allowing $\theta_h \in (0, 0.5)$.

Prior to variant analysis, knowledge about θ_t is usually not available, because only variant analysis itself can yield insight into the clonal structure of a tumor. It is therefore reasonable to assume that $\theta_t \in [0, 1]$, that is we allow any possible VAF to apply for a somatic variant in a tumor cell.

A tumor genome sample can still contain non-negligible amounts of healthy cells, which affects our considerations. Therefore, let $\alpha \in [0, 1]$ be the *purity* of the tumor genome sample, that is the relative amount of fragments in the sample that stem from a cancerous genome copy. Thus, $1 - \alpha$ is the relative amount of fragments that stem from a healthy genome copy from within the tumor genome sample. For θ_c, defined to be the VAF among only the cancer cells in the tumor sample, we obtain the relationship

$$\theta_t = \alpha \cdot \theta_c + (1 - \alpha) \cdot \theta_h \tag{7}$$

which, just as θ_t can take values in all of the unit interval $[0, 1]$ for a somatic variant.

Finally, *strand bias* tends to introduce non-negligible issues in variant analysis. Our analyses are affected by strand bias as well. Strand bias refers to the fact that during the sequencing process, certain sequence motifs can cause the sequencing process to slow down, or even temporarily stall. If such delays occur, systematic sequencing artifacts affect the reads. Because the motif disappears when considering the complementary strand (unless the motif is a reverse complementary palindrome), artifact-inducing motifs usually affect reads from only one of the strands. This implies that the occurrence of artifacts depends on the origin of the read: artifacts can only affect reads from one of the strands, either the forward or the reverse strand, while reads sampled from the other strand are not affected by the sequencing artifact. See Allhoff et al. [2013] for more details.

In summary, there are three possible cases we need to deal with: *first*, a putative variant affects only reads sampled from the reverse strand, *second*, the putative variant affects reads from forward and reverse strands at equal rates, or *third*, the putative variant only affects reads sampled from the forward strand. The exclusive support of a variant by reads from only one of the two strands is indicative of an artifact [Li, 2014]. The goal is to remove such artifacts from the output.

We approach this by introducing a variable β taking values in $\{0, \frac{1}{2}, 1\}$ reflecting the three cases from above (so $\beta = 1$ reflects that the variant only shows on reads that stem from the forward strand, and so on). In other words, β reflects the probability that a read that is associated with the variant stems from the forward strand.

5.2 Motivation: Why Naive Approaches to Computing (1) Fail.

To understand why *efficient computation of* (1) *is difficult*, consider that each of the reads Z_i^h, Z_j^t could

(a) not stem from the particular variant locus,

(b) stem from the locus, but is not affected by the variant,

(c) stem from the locus, and is indeed affected by the variant.

We recall that it can be particularly difficult to be certain about (a), (b) or (c) when dealing with reads being associated with midsize indel loci (30-250 bp; sometimes termed the ”NGS twilight zone”). Let $k = |\boldsymbol{Z}^t|$ and $l = |\boldsymbol{Z}^h|$ be the read coverage of the locus in the tumor and the healthy sample. Since there are 3 different possibilities—namely (a), (b) or (c)—for the overall $k + l$ reads, we obtain that there are 3^{k+l} different scenarios that could reflect the truth, all of which apply with a particular probability. For computing

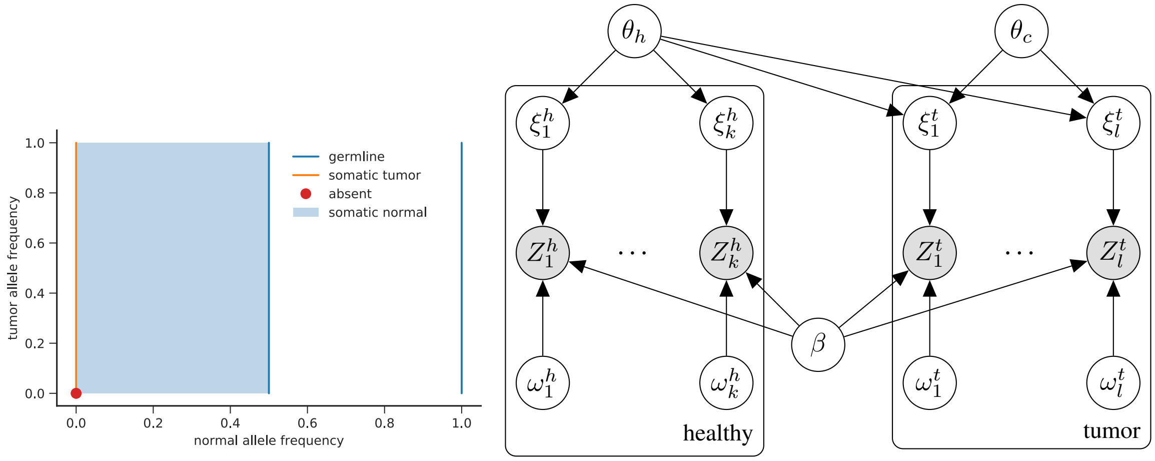


Figure 8: Left: Visualization of the parameter space Θ of the VAFs. *Orange*: somatic variants agree with $(\theta_h, \theta_c) \in \{0\} \times (0, 1]$, which means that no healthy cells have the variant ($\theta_h = 0$), while some cancer clones do have the variant ($\theta_c > 0$). Germline variants (*blue*) are described by $\theta_h \in \{\frac{1}{2}, 1\}$ and absent variants (*red dot*) by $\theta_h = 0, \theta_c = 0$. Subclonal somatic variants in the healthy (normal) tissue are described by $\theta_h \in (0, 0.5)$. Right: Diagram of the model presented in section 5.3 (white circles=latent variables; grey circles=observable variables). Each column corresponds to one alignment (Z_i^h or Z_j^t) with its hyperparameters ξ_i^h, ω_i^h or ξ_j^t, ω_j^t. Due to (potential) sample impurity (denoted by α in the text), θ_h has an influence on the alignments Z_j^t from the tumor sample.

(1) following a *fully Bayesian approach to inverse uncertainty quantification* [Liu et al., 2009]—which is the approved and canonical way to quantify uncertainties in our setting—one needs to integrate over all the possible $k + l$ choices. In a naive approach, this translates into computing a sum with 3^{k+l} summands. Because $k + l$ amounts to at least 60 to 70 in standard settings, naive approaches fail to compute the integral in human feasible runtime. This is further aggravated because one usually needs to consider hundreds of thousands of putative indel loci. *So, methodical efforts are required for uncertainty quantification in our setting.*

5.3 The Model

We present a graphical model that captures all dependency relationships among the variables relating to the computation of $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ while taking all major uncertainties into account (see Figure 8 **(b)**).

Beyond observable variables reflecting observable read data $Z_i^h, i = 1, ..., k, Z_j^t, j = 1, ..., l$ and latent variables $\theta_h, \theta_c, \beta$ for allele frequencies and strand bias, we first introduce two additional latent variables. We denote with $\xi_i^h, \xi_j^t \in \{0, 1\}$ whether fragments Z_i^h, Z_j^t are associated with the variant ($\xi_i^h, \xi_j^t = 1$) or not ($\xi_i^h, \xi_j^t = 0$). Further, $\omega_i^h, \omega_j^t \in \{0, 1\}$ indicate whether reads indeed stem from the locus ($\omega_i^h, \omega_j^t = 1$) or not ($\omega_i^h, \omega_j^t = 0$).

Variables $\omega_i^h, \omega_j^t, \xi_i^h, \xi_j^t$ refer to the above-mentioned three cases (a), (b), (c). For example $\xi_i^h = 1, \omega_i^h = 1$ represents the case that read Z_i^h stems from the locus of interest and and is indeed associated with the variant (case (c) above). Note that the case $\omega = 0$, which indicates that the read does not stem from the locus (a), renders specification of other variables obsolete, because this particular read cannot provide information about the variant. Since knowledge about realizations of the hyperparameters cannot be observed at the time of the analysis, they are latent variables.

In the following, we introduce the full model in three steps, thereby deriving the conditional dependencies between the different variables. *First*, we consider the basic model. *Second*, we introduce *impurity*, modeling the fact that the cancer genome sample also contains healthy genome copies. This implies to distinguish between the treatment of reads from healthy and tumor sample: while reads from the healthy sample still follow the basic model, processing reads from the tumor sample requires a modification to take impurity into account. *Third*, we introduce *strand bias*, which would affect both reads from the healthy and the tumor sample and thus requires to apply modifications for both parts.

Part 1 – Foundations: Alignment and Typing Uncertainty. We first model the two major sources of uncertainty, (a) *alignment uncertainty* and (b) *typing uncertainty*. Let $Z_i, i = 1, \ldots, k$ be the observable alignment data. *Alignment uncertainty* is handled via

$$\omega_i \sim \text{Bernoulli}\,(\pi_i) \tag{8}$$

where π_i reflects the probability that the i-th read has been aligned to the correct position in the genome. Estimates for π_i are provided by the aligner of choice, via the reported MAPQ values [Li et al., 2008a]. *Typing uncertainty* can be modeled as

$$\xi_i \sim \text{Bernoulli}\,(\theta\tau) \tag{9}$$

Thereby, the allele frequency θ, as per its definition, reflects the probability to sample a read from a variant-affected genome copy. Further, τ reflects the probability that, if sampled from the variant-affected copy, the read indeed covers the variant. That is, the product $\theta\tau$ reflects to sample a fragment that is truly affected with the variant (and hence provides real evidence about the variant). See section 5.3.2 for how τ is computed. Note that θ and τ vary depending on whether Z_i refers to fragment data from the healthy (Z_i^h) or the tumor genome (Z_j^t). Appropriate choices for θ, τ are specified in the paragraph about impurity below.

Whether $\xi_i = 1$ or $\xi_i = 0$ is generally not immediately evident from the observed (paired-end) read Z_i, leading to *typing uncertainty*. We define

$$Z_i \mid \xi_i, \omega_i \sim \begin{cases} p_i & \text{if } \xi_i = 1, \omega_i = 1 \\ a_i & \text{if } \xi_i = 0, \omega_i = 1 \\ o_i & \text{if } \omega_i = 0 \end{cases} \tag{10}$$

where a_i, p_i, o_i are the probability distributions that reflect the situation that Z_i stems from a genome copy in which the indel is either *p*resent or *a*bsent, or comes from an (unknown) *o*ther locus. In the last case, Z_i is supposed to have no influence on the posterior probability distribution of θ.

In order to illustrate the nature of p_i, a_i, o_i, let us consider a simplified case. The detailed definition can be found in section 5.3.1. We denote two different haplotypes H_{ref} and H_{var}. The former represents the reference sequence (no variant) and the latter the alternative (variant affected) sequence at the considered locus. Let x be a particular read. Let further

$$P_{\text{ref}}(x) \quad \text{and} \quad P_{\text{var}}(x) \tag{11}$$

be the probabilities that x has been sampled from H_{ref} or from H_{var}, respectively. Because x may contain errors, one needs to take the sequencing error profile of x into account when aiming at accurate computation of $P_{\text{ref}}(x)$ and $P_{\text{var}}(x)$. The error profile is provided via the base qualities reported by the sequencing machine[21].

[21] Base qualities are reported along with read alignments in BAM files, see `https://samtools.github.io/hts-specs/SAMv1.pdf`.

As described in earlier work [Poplin et al., 2018], probabilities (11) can be reliably computed by means of a Pair HMM whose parameters refer to the base quality profile of x, thereby appropriately accounting for the sequencing errors affecting x. Following this well-approved rationale, we model

$$a_i(Z_i = x) \overset{\text{def}}{=} P_{\text{ref}}(x) \quad \text{and} \quad p_i(Z_i = x) \overset{\text{def}}{=} P_{\text{var}}(x) \tag{12}$$

In a final remark, note that for that sake of a clear presentation, we have omitted the detail that p_i, a_i also reflect considerations about the fragment length distributions of the involved reads. See Subsection 5.3.1 for how a_i, p_i and o_i are computed in full detail.

Finally, o_i reflects the case that x does not stem from the locus. Without further knowledge available—which is the case at the time of analysis—it is reasonable to assume that o_i is equal for all possible reads x. In other words, it is reasonable to assume that o_i reflects a uniform distribution. The only detail to consider (which follows from the theoretical statements listed in section 5.3.4), is that the particular value $o(Z_i = x)$ needs to scale right relative to $p_i(Z_i = x)$ and $a_i(Z_i = x)$, see again section 5.3.1 for the respective details.

Part 2 – Impurity. Let α be the purity of the tumor genome sample. In turn, $1-\alpha$ is the ratio of fragments stemming from a healthy genome copy, when sampled from the tumor sample. Impurity does not affect the healthy sample, so (8), (9), (10) and (12) apply without further modifications for the healthy sample: variables can be indexed with super- or subscript h as they appear. For example $\omega_i^h \sim \text{Bernoulli}(\pi_i^h)$ in (8) or $\xi_i^h \sim \text{Bernoulli}(\theta_h \tau_h)$ in (9).

However, variables referring to the tumor sample (indexed using super- or subscript t) require different treatment. Let $Z_j^t, j = 1, \ldots, k$ be the observable read data in the tumor sample. First note that impurity does not affect the degree of certainty of read alignments. So (8) applies without modifications:

$$\omega_j^t \sim \text{Bernoulli}\left(\pi_j^t\right) \tag{13}$$

Modeling typing uncertainty however requires changes. Recalling (7), we compute

$$\xi_j^t \sim \text{Bernoulli}\left(\theta_t \tau_t\right) \overset{(7)}{=} \text{Bernoulli}\left(\alpha \theta_c \tau_t + (1-\alpha)\theta_h \tau_t\right) \tag{14}$$

This reflects that if reads stem from a cancerous genome copy, which happens with probability α, variables θ_c, τ_t apply, and if stemming from a healthy genome copy, which happens with probability $1-\alpha$, variables θ_h, τ_t apply. Note that although stemming from a healthy genome copy (thus θ_h), still τ_t applies, because the healthy genome copy was sampled from the tumor sample and τ is specific to the sample, healthy or tumor (see section 5.3.2 for details).

Equations (10) and (12) remain unchanged, that is super- and subscripts t can be introduced without further modifications, because impurity does not matter in these considerations.

Part 3 – Strand Bias. In the following, we will be dealing with paired-end read data. We will therefore be focusing on this case. Note however that our model also immediately applies for single-end and mate-pair read data, via some straightforward modifications.

For sake of simplicity, let us consider reads from only the healthy sample, and for the sake of a clear presentation, we omit super- and subscripts. That is, we write $\mathbf{Z} = (Z_i)_{i=1}^k$ when meaning $\mathbf{Z}^h = (Z_i^h)_{i=1}^k$ and we simply write θ, τ when meaning θ_h, τ_h. Extending the following arguments to the tumor sample is straightforward.

Following the usual protocols, in a paired-end read, one read end stems from the forward strand, while the other end stems from the reverse strand. When accounting for strand bias, it is important to track whether the forward or the reverse strand is associated with the variant, or even both of them (reflecting the situation of a self-overlapping read), because variants showing exclusively in only one of the types of ends are likely to reflect artifacts.

We model strand direction of read pairs by expanding $\mathbf{Z}$ into $\mathbf{Z} = (\mathbf{R}, \mathbf{S})$, distinguishing between the read data themselves $\mathbf{R} = (R_i)_{i=1}^k$ and variables $\mathbf{S} = (S_i)_{i=1}^k$. Each of the S_i takes values in $\{-, +\}$, reflecting whether, the reverse $(-)$ or the forward $(+)$ end of read pair i are associated with the variant. For the sake of a clear presentation, we omit the case that a paired-end read has self-overlapping ends, such that it is possible that both of its ends are associated with the variant; see section 5.3.3 for the corresponding, straightforward modifications.

The value of S_i only has meaning if the i-th read pair indeed stems from the variant haplotype, which refers to the case $\xi_i = 1, \omega_i = 1$ in terms of the earlier latent variables. If $\omega_i = 0$, that is the i-th read pair does not stem from the variant locus, or if $\omega_i = 1, \xi_i = 0$, that is, if the i-th read pair stems from the locus but is associated with the reference haplotype, strand bias cannot affect the i-th read pair. Realizations of S_i are observable: we retrieve the corresponding values from the initial standard read aligner in the obvious way. Note that it is important to retrieve the realizations from the standard alignment, but not the re-alignment, because standard, reference-based alignments are a major source of strand bias artifacts.

Integrating the S_i, the likelihood function extends to

$$L(\theta, \beta \mid \mathbf{Z} = (\mathbf{R}, \mathbf{S})) \propto P(\mathbf{Z} = (\mathbf{R}, \mathbf{S}) \mid \theta, \beta) = \prod_{i=1}^{k} P(R_i, S_i \mid \theta, \beta)$$

So, when treating strand bias we have to specify how to evaluate $P(R_i, S_i \mid \theta, \beta)$. We compute

$$P(S_i, R_i \mid \theta, \beta) = P(S_i \mid R_i, \theta, \beta) \times P(R_i \mid \theta, \beta) = P(S_i \mid R_i, \theta, \beta) \times P(R_i \mid \theta) \tag{15}$$

where the last equality follows from the fact that only S_i depends on β, while R_i, the observed read sequence itself, does not depend on β. Note that R_i could be identical both for $(+, -)$-oriented and $(-, +)$-oriented read pairs, while only one of those orientations gives rise to a variant artifact. Computing $P(R_i \mid \theta)$ is identical with the computations displayed for the cases not involving strand bias, see Part 1 and 2. So it remains to consider $P(S_i \mid R_i, \theta, \beta)$. We refer to section 5.3.3 where we elaborate on the corresponding details; here we conclude that (15) points out that strand bias can be handled efficiently by integrating additional factors $P(S_i \mid R_i, \theta, \beta)$ into the overall likelihood function.

As further outlined in the last paragraph of section 5.3.3, the factor $P(S_i \mid R_i, \theta, \beta)$ in (15) has no influence on the likelihood of θ if $\beta = \frac{1}{2}$. This is important, because only loci where $\beta = 0$ or 1 reflect strand bias artifacts and are to be removed from further considerations. Variants observed at loci where $\beta = \frac{1}{2}$ are no strand bias artifacts. Therefore, the desirable scenario is to deal with them as if strand bias was not to take into account. This means that $L(\theta, \frac{1}{2} \mid \mathbf{Z} = (\mathbf{R}, \mathbf{S}))$ should be proportional to $L(\theta \mid \mathbf{Z})$, as treated before introducing strand bias; see again section 5.3.3 for further straightforward computations that prove this.

5.3.1 Computing a_i, p_i and o_i

We now specify distributions a_i and p_i for read pairs Z_i. We note in passing that our model does in general not depend on a particular sequencing technology, so can also deal with single-end third-generation sequencing reads or other protocols; see the Discussion for some final remarks on that point.

As usual, let us fix a particular locus harboring a putative variant and denote the corresponding reference haplotype of that locus by H_{ref} and the alternative, variant-affected haplotype of the locus by H_{var}. We compute H_{var} by application of the putative variant to the reference sequence (H_{ref}) without any additional changes.

Let us consider a particular read pair $x = (x_1, x_2)$, consisting of two ends x_1, x_2 (which in general are of equal length), that was found to align with the putative variant locus through use of a standard read aligner, (e.g. BWA-MEM Li [2013]). After having collected x, we discard all standard read alignment information about x, apart from one detail: we store z, the length of the alignment computed by the standard read aligner. Note that z evaluates as the distance between the rightmost alignment coordinate of the right end (x_2) and the leftmost alignment coordinate of the left end (x_1), including clipped bases.

Subsequently, we re-align x with both H_{ref} and H_{var} using a Pair HMM [Durbin et al., 1998], as it was initially suggested by Poplin et al. [2018]. The Pair HMM appropriately accounts for possible sequencing errors in x, because its parameters reflect the sequencing error (base quality) profile of x, and thereby reflects that x has been sampled as a fragment from H_{ref} or H_{var}, by providing either x and H_{ref} or x and H_{var} as input pair of sequences.

We base all of our further considerations on the resulting re-alignments of x with H_{ref} and H_{var}. The well-known justification for this is that such re-alignments avoid several biases standard read alignments (where only H_{ref} is considered) are affected with, and are therefore of considerably higher quality. When dealing with insertions and deletions, these biases include well-known effects referring to mistaken gap placement—note that placing gaps in alignments correctly has remained a notorious issue [Lunter et al., 2008]. Incomplete gaps (soft-/hard-clipped read alignments) output by the standard aligner add to these difficulties; therefore re-aligning clipped alignments can be of particular value. See Figure 9a for an illustration and Figure S8 for a read world example.

We derive values $a_i(x), p_i(x), o_i(x)$ from these re-alignments, reflecting probabilities that x has been sampled from H_{ref} (a_i), or from H_{var} (p_i), while o_i, reflecting the probability that x does not stem from the locus, needs to scale right with a_i and p_i, in order to ensure numerical stability.

For aligning x with H_{ref} and H_{var} using a Pair HMM, we need to consider the issue that we need to specify exactly what $H_{\text{ref}}, H_{\text{var}}$ should look like. Of course, $H_{\text{ref}}, H_{\text{var}}$ cannot reflect whole-chromosome length sequences, because it is infeasible by runtime to align x against a full-length chromosome. Hence, we need to cut H_{ref} out of the full-length chromosome in an appropriate way. In particular, it is favorable that the parts of H_{ref} and H_{var} against which x is aligned (i.e., the parts provided as input to the Pair HMM) should be of equal length (or differ by one or two basepairs, but not differ by the length of the indel, for example), to avoid biases induced by the length of the indel under consideration.

To do this, we proceed in six steps:

1. Let l be the coordinate of the variant locus in the reference genome G. We determine $H_{\text{ref}} := G[l - h, l + h]$, that is as a window of size $2h$ in the reference genome around the variant locus[22]. We then obtain H_{var} by applying the variant to H_{ref}. Note that this is not yet the input for the Pair HMM.

2. Let x_1 be the left end of the read pair $x = (x_1, x_2)$. The following computations are executed also for x_2, which happens entirely analogously. Taking x_1 as the pattern and H_{ref} or H_{var}, respectively, as the text, we apply Myers' bitvector algorithm [Myers, 1999] and obtain coordinates a_r, b_r and a_v, b_v determined such that, relative to edit distance, the optimal occurrence of x_1 in H_{ref} and H_{var} is $H_{\text{ref}}[a_r, b_r]$ and $H_{\text{var}}[a_v, b_v]$, respectively. Note that by virtue of Myers' bitvector algorithm, the lengths of $H_{\text{ref}}[a_r, b_r]$ and $H_{\text{var}}[a_v, b_v]$, that is, $b_r - a_r + 1$ and $b_v - a_v + 1$, tend to differ only by at most 1 or two basepairs, which is the desired scenario. As above-mentioned, we also obtain such coordinates for the right end x_2.

3. We finally specify a maximal edit distance d, and compute

$$a_{i,\text{left}}(x) := P_{\text{HMM}}(x_1, H_{\text{ref}}[a_r - d, b_r + d]) \tag{16}$$

$$p_{i,\text{left}}(x) := P_{\text{HMM}}(x_1, H_{\text{var}}[a_v - d, b_v + d]) \tag{17}$$

[22] Varlociraptor uses $h = 64$ by default.

where $P_{\text{HMM}}(x_1, H_{\text{ref}}[a_r - d, b_r + d])$ and $P_{\text{HMM}}(x_1, H_{\text{var}}[a_v - d, b_v + d])$ are computed by means of the forward algorithm for Pair HMMs as defined by Durbin et al. [1998]. Again, probabilities $a_{i,\text{right}}(x), p_{i,\text{right}}(x)$ referring to the right read end x_2 are computed in the entirely analogous way, by replacing x_1 with x_2 and coordinates a_r, b_r, a_v, b_v with coordinates resulting from running Myers' algorithm with x_2 as pattern and H_{ref} and H_{var} as texts.

4. Let f specify the fragment length distribution referring to the sequencing library protocol in use. Note that f often is approximately Gaussian, hence can be specified by its mean μ and standard deviation σ. While, without any further adaptations, we can work with arbitrary (empirical) fragment length distributions, we make use of Gaussian distributions here.

   We recall that z is the length of the (standard) alignment of x. Note that z agrees with the length of the underlying fragment if x is not affected by the variant. If, however, x is affected by the variant, which is of length δ, then the length of the underlying fragment would evaluate as $z + \delta$. Hence, we define

$$a_{i,\text{int}}(x) := f(z) \tag{18}$$

$$p_{i,\text{int}}(x) := f(z + \delta) \tag{19}$$

   to reflect that $a_{i,\text{int}}$ is supposed to reflect that alignment and fragment length agree, while $p_{i,\text{int}}$ is supposed to reflect that they differ by the length of the variant.

5. We combine the evidence for the left and right read with the insert size by calculating

$$a_i(x) := a_{i,\text{left}}(x) \times a_{i,\text{int}}(x) \times a_{i,\text{right}}(x) \tag{20}$$

$$p_i(x) := p_{i,\text{left}}(x) \times p_{i,\text{int}}(x) \times p_{i,\text{right}}(x). \tag{21}$$

6. Finally, we denote

$$o_i(x) := \frac{1}{2}(a_i(x) + p_i(x)). \tag{22}$$

   The formal justification for determining $o_i(x)$ as such follows from the proofs provided in section 5.3.4, which lists the formal statements that make the theoretical foundation of our model.

Note that we discard all reads from being considered if neither the left read end (x_1) nor the right read end (x_2) overlap the variant locus as per their Pair HMM based re-alignments.

Finally, let us revisit the decision to combine evidences of left and right read with the insert size. Splitting up of a_i, p_i into read end-specific factors would effectively weaken variant related signals from reads that can only be ambiguously placed, that is whose exact placement remains dubious even after re-alignment. To understand the advantage of combining evidences, consider a read x whose first end supports the variant, which yields large $p_{i,\text{left}}(x)$, but small $a_{i,\text{left}}(x)$, whereas the second end supports the reference, which yields large $a_{i,\text{right}}(x)$, but small $p_{i,\text{right}}(x)$. In consequence, overall, both $a_i(x)$ and $p_i(x)$ are approximately equal. The same argument holds, for example, if one or both reads support the variant, but the insert size does not. As follows from our model, reads x with (approximately) equal $a_i(x)$ and $p_i(x)$ yield an approximately uniform probability distribution with respect to θ, the allele frequency of the variant, which is just our prior distribution on θ. In other words, x does not contribute to making a statement in favor of a particular (range of) θ, which is the correct scenario.

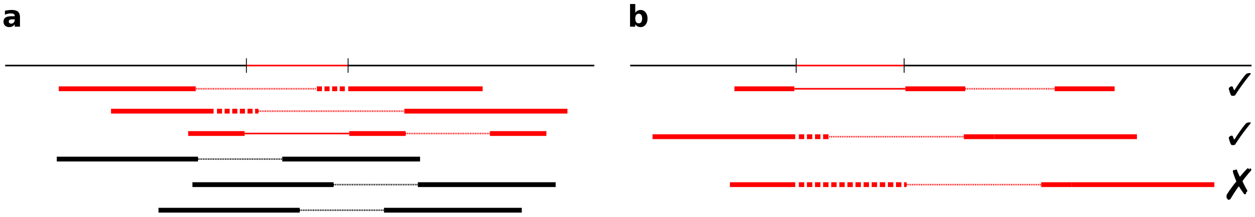


Figure 9: Dealing with evidence from alignments. The reference genome is displayed as a thin line at the top, with an example deletion highlighted in red. Paired-end reads are displayed as two connected thick bars. Deletions within the alignment of a read are displayed as thinner region. Soft-clips in the alignment are dotted. Fragment origin is encoded by color (red=variant; black=reference). (a) Reads coming from the variant allele can be aligned with soft-clipped ends, or with the variant being encoded in the alignment itself. Due to ambiguities, the positions are not necessarily perfectly aligned with the variant (see Figure S8 for a real example). (b) Depending on the size of the variant and properties of the read mapper on a specific genome, it can become less likely to obtain fragments from the variant allele. Here, the first and the second fragment are mappable, the third is not, because the soft-clip would become too large to be considered by the read mapper.

5.3.2 Sampling Probability

See Figure 9b for illustrations of the following. In the initial step, we select reads using a standard read aligner, which determines placement of reads by mapping them against the reference genome. Because differences between non-variant-affected reads and the reference sequence are small, the standard read aligner will be able to align all (or at least nearly all) of such reads successfully against the putative variant locus. The situation is different for reads that are affected by the variant. Because such reads differ from the reference sequence by, in our case, an insertion or deletion of considerable length, the standard read aligner may fail to align such reads successfully. Such failure occurs in particular if variant size and placement within the read interfere unfavorably with the alignment procedure. As a consequence, there can be a bias towards reads from the reference allele.

The effect was first systematically treated in [Sahlin et al., 2015]. The corresponding quantification means to assume that non-variant affected reads align against the locus with probability 1, while variant-affected reads align with *sampling probability* τ, which is smaller than one. The sampling probability τ is a locus-, donor genome-, and sequencing-specific parameter, because τ depends on the sequence context, which varies relative to the locus considered and also relative to the donor genome investigated as well as the used sequencing protocol (in terms of read length and targeted insert size). So, τ can be different for cancer and control genome, which we make explicit by dealing with τ_h and τ_t, referring to the healthy (h) and tumor (t) genome.

In the following, we briefly specify how τ is computed based on the considerations made by Sahlin et al. [2015]. First, τ depends on the fragment length distribution f, which can be retrieved from the read aligner; see also explanations referring to formulas (18) and (19) above. Further, τ depends on the number of bases δ the deletion or insertion covers in the donor genome (deletion: $\delta = 0$; insertion: $\delta =$ length of insertion). It also depends on the number of bases the aligner requires to be aligned with the reference genome sequence upstream or downstream of the variant, in the upstream (left) or the downstream (right) read end, respectively, which are referred to as k_{up} and k_{down}. This reflects that it is critical for a read aligner to get the outer ends of a paired-end read mapped. Values k_{up} and k_{down} depend on the read aligner and the variant. In the following, k is one of k_{up} or k_{down}, depending on whether we consider information referring to the upstream or downstream read end; when dealing with single-end reads (see below), k is unique. k can be estimated by inspecting a representative subset of all aligned reads of a sample, and recording the maximum deletion and insertion size encoded in the CIGAR strings [Li et al., 2009] of this subset of reads, as well as the maximum soft clip size, retrieved from partial alignments from that representative subset of reads. If the variant is smaller than the maximum deletion or insertion encoded in the CIGAR strings, we can set $k = 0$ because the variant is small enough to be part of the alignment. Otherwise, k can be set to the read length minus the maximum soft clip size, that is, the smallest observed partial alignment.

Let further o denote the fragment length (which is unknown) and f denote the fragment length distribution (see section 5.3.1). Following Sahlin et al. [2015], we compute τ as

$$\tau_s = \sum_{o=1}^{\infty} f(o) \frac{o - \delta - k_{\text{up}} - k_{\text{down}}}{o - \delta}, \tag{23}$$

where $s = h, t$ specifies the healthy or tumor sample, see above; note that all of f, k_{up} and k_{down} differ relative to the sample. The formula reflects that one sums over all possible fragment lengths, weighted by the probability to sample a fragment of that length, and calculates the probability to get a fragment of that length aligned, relative to the relevant parameters δ, k_{up} and k_{down}. Note that in case of single end reads, (23) simplifies to

$$\tau_s = \frac{r - \delta - k}{o - \delta}, \tag{24}$$

with r being the read length.

5.3.3 Strand Bias: Technical Details.

In the following, observable strand bias variables S_i will take values in $\{-, +, \pm\}$ to also reflect the case that a paired-end read has self-overlapping reads both of which are affected by the variant ($\pm$). Note that existence of a $\pm$-read rules out strand bias, because it means that read ends from both strands are affected by the variant.

Let us define

$$\begin{aligned} A &:= \{\omega_i = 1, \xi_i = 1\} && (25)\\ B &:= \{\omega_i = 1, \xi_i = 0\} && (26)\\ C &:= \{\omega_i = 0\} && (27)\\ D &:= B \mathbin{\dot{\cup}} C && (28) \end{aligned}$$

and recall that strand bias can only occur if event A applies, that is, if the read is indeed associated with the variant. Note that A and D span the entire space of cases $(\omega_i, \xi_i) \in \{0, 1\} \times \{0, 1\}$, which allows to apply the law of total probability. We obtain

$$P(S_i \mid Z_i, \theta, \beta) = P(S_i, A \mid Z_i, \theta, \beta) + P(S_i, D \mid Z_i, \theta, \beta) \tag{29}$$

So we will continue to put further focus on the two summands in (29). Let z_i be the observed realization (the read itself including its error profile) of Z_i. For a clear presentation of what follows we define

$$\gamma_A(Z_i) := p_i(z_i)\pi_i\theta\tau \quad \text{and} \quad \gamma_D(Z_i) = a_i(z_i)(\pi_i(1 - \theta\tau)) + o_i(z_i)(1 - \pi_i). \tag{30}$$

For the first summand from (29), we compute

$$\begin{aligned} P(S_i, A \mid Z_i, \theta, \beta) &= P(S_i \mid A, Z_i, \theta, \beta) \cdot P(A \mid Z_i, \theta, \beta) \\ &= P(S_i \mid A, \beta) \cdot P(A \mid Z_i, \theta) \end{aligned} \tag{31}$$

where the last equation follows from the fact that, on the one hand, given A and β, S_i is independent of Z_i and θ (note that all information S_i depends on about Z and θ is captured by A) and, on the other hand, A is independent of β, the strand bias. We continue

$$\begin{aligned} P(A \mid Z_i, \theta) &= \frac{P(A, Z_i \mid \theta)}{P(Z_i)} = \frac{1}{P(Z_i)} P(Z_i \mid A, \theta) P(A \mid \theta) \\ &= \frac{1}{P(Z_i)} P(Z_i \mid A) P(A \mid \theta) = \frac{1}{P(Z_i)} \gamma_A(Z_i) \end{aligned}$$

where the second to last equation follows from the dependency structure of the model, because Z_i is independent of θ given A, and the last equation follows from the fundamental considerations of Part 1. Note at last that it is reasonable (and common in analogous settings) to assume that $P(Z_i)$, the prior probability to observe a read is constant across all reads, so $\frac{1}{P(Z_i)}$ turns out to be a constant. In summary, we obtain

$$P(S_i, A \mid Z_i, \theta, \beta) \propto \gamma_A(Z_i) \cdot P(S_i \mid A, \beta). \tag{32}$$

Recalling that read pairs were selected where at least one of the ends overlapped the locus, let q_1 and $q_2 = 1 - q_1$ be the probabilities to have one or both reads overlapping the variant locus. Probabilities q_1, q_2 can be determined from the insert size distribution provided by the read mapper, in combination with the length of the read ends. Note that usually $q_1 >> q_2$, and that for single-end sequencing, obviously $q_1 = 1, q_2 = 0$. Our considerations are then finalized by providing the table

$$P(S_i \mid A, \beta) = \begin{cases} 1 & \text{if } \begin{cases} S_i = +, \beta = 1 \\ S_i = -, \beta = 0 \end{cases} \\ \frac{1}{2} q_1 & \text{if } \quad S_i \neq \pm, \beta = \frac{1}{2} \\ q_2 & \text{if } \quad S_i = \pm, \beta = \frac{1}{2} \\ 0 & \text{if } \begin{cases} S_i = +, \beta = 0 \\ S_i = -, \beta = 1 \\ S_i = \pm, \beta \neq \frac{1}{2}. \end{cases} \end{cases} \tag{33}$$

By entirely analogous considerations, we also obtain

$$P(S_i, D \mid Z_i, \theta, \beta) \propto \gamma_D(Z_i) \cdot P(S_i \mid D, \beta) \tag{34}$$

where here, however,

$$P(S_i \mid D, \beta) = \frac{1}{3} \quad \text{for all combinations of realizations of } S_i, \beta \tag{35}$$

As an exemplary case, consider $\gamma_A(Z_i) = 0, \gamma_D(Z_i) = 1$, reflecting the case that the read is not associated with the variant with probability one. Here, we obtain that $P(S_i \mid Z_i, \theta, \beta) \propto \frac{1}{3}\gamma_D(Z_i)$ for all choices of β. This means that the i-th read assigns equal likelihood to all choices of β, which is the correct scenario, because the i-th read cannot provide any information about β.

Consider also the opposite case, $\gamma_A(Z_i) = 1, \gamma_D(Z_i) = 0$, reflecting full certainty about the read being associated with the variant. If for example $S_i = +$, the likelihood for $\beta = 0$ is zero. Further (recalling that usually $q_1 >> q_2$, meaning that q_1 is close to one), the likelihood for $\beta = 1$ is approximately double the amount as for $\beta = \frac{1}{2}$. This again is the correct scenario.

Finally, we observe that if $\beta = \frac{1}{2}$, the factor $P(S_i \mid Z_i, \theta, \beta)$ in (15) has no influence on the likelihood of θ. This is important, because we will only consider loci where $\beta = \frac{1}{2}$ with large enough likelihood, and for such loci, we will base our further decisions on the evaluation of $L(\theta, \frac{1}{2} \mid \mathbf{Z}, \mathbf{S})$. It is therefore essential that $L(\theta, \frac{1}{2} \mid \mathbf{Z}, \mathbf{S})$ is proportional to $L(\theta \mid \mathbf{Z})$, that is when not considering strand bias. Considering strand bias should only lead to excluding variant artifacts—if variants are supposed to be real, strand bias considerations should not have an influence on decisions about whether variants are somatic or not, or, more specifically, about what their allele frequencies are.

To understand this, note that the only appearance of θ in the computation of $P(S_i \mid Z_i, \theta, \beta)$ is in the factors $\gamma_A(Z_i)$ and $\gamma_D(Z_i)$. Note further that these factors determine to what degree the i-th read makes a statement about β: the larger $\gamma_A(Z_i)$ relative to $\gamma_D(Z_i)$, the stronger the statement. If $\beta = \frac{1}{2}$, however, both $P(S_i \mid A, \beta)$ and $P(S_i \mid D, \beta)$ evaluate as $\frac{1}{2}$, independently of the particular realization of S_i. Therefore, if $\beta = \frac{1}{2}$, varying θ does not vary $P(S_i \mid Z_i, \theta, \beta)$. Of course, still, varying θ varies (15) because it varies $P(Z_i \mid \theta)$. So, θ varies $L(\theta, \frac{1}{2} \mid \mathbf{Z}, \mathbf{S})$, just as if β was not considered. In summary, this is the desired scenario.

5.3.4 Statements.

We finally obtain the result outlined in section 2.2 as a corollary to the following theorem.

Theorem 5.1. *Let $\boldsymbol{Z}^h = (Z^h_1, ..., Z^h_k), \boldsymbol{Z}^t = (Z^t_1, ..., Z^t_l)$ be the observable read data from a healthy and a tumor sample, covering the locus of a putative variant. Then*

- (i) *The likelihood function*

$$L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t) = \prod_{i=1}^{k} L(\theta_h, \theta_c, \beta \mid Z^h_i) \times \prod_{j=1}^{l} L(\theta_h, \theta_c, \beta \mid Z^t_j) \tag{36}$$

factors into likelihood functions referring to individual read pairs.

- (ii) *Let Z_i refer to any of the read data $Z^h_1, ..., Z^h_k, Z^t_1, ..., Z^t_l$ and let ω_i, ξ_i be its latent uncertainty hyperparameters. Then*

$$\begin{aligned} L(\theta_h, \theta_c, \beta \mid Z_i) &= P(Z_i \mid \theta_h, \theta_c, \beta) \\ &= \int_{\xi_i, \omega_i} P(Z_i \mid \xi_i, \omega_i) \times P(\xi_i, \omega_i \mid \theta_h, \theta_c, \beta)\ d(\xi_i, \omega_i) \end{aligned} \tag{37}$$

PROOF. (i) follows immediately from the fact that the Z^h_i, Z^t_j are conditionally independent given $\theta_h, \theta_c, \beta$; see Figure 8b. (ii) follows from application of the Chapman-Kolmogorov equation, in combination with the dependency relationships captured by our model, see again Figure 8b. □

We distinguish between read data Z^h_i from the healthy sample and read data Z^t_j from the tumor sample. In the following, we make use of the fact that ω^h_i and θ_h are independent (see Figure 8b). Let $s^h_i = P(S^h_i | \omega^h_i = 1, \xi^h_i = 1, \beta)$ be the likelihood of the strand bias given read pair i as defined in equation (33).

We compute the likelihood function for $Z_i^h = (R_i^h, S_i^h)$, which does not depend on θ_c, as

$$\begin{aligned}
P(Z_i^h \mid \theta_h, \theta_c, \beta) &= \int_{\xi_i^h, \omega_i^h} P(Z_i^h \mid \xi_i^h, \omega_i^h, \beta) \times P(\xi_i^h, \omega_i^h \mid \theta_h, \theta_c)\, d(\xi_i^h, \omega_i^h) \\
&= \int_{\xi_i^h, \omega_i^h} P(Z_i^h \mid \xi_i^h, \omega_i^h, \beta) \times P(\xi_i^h, \omega_i^h \mid \theta_h)\, d(\xi_i^h, \omega_i^h) \\
&= P(Z_i^h \mid \omega_i^h = 1, \xi_i^h = 1, \beta) P(\omega_i^h = 1, \xi_i^h = 1 \mid \theta_h, \beta) + \\
&\quad P(Z_i^h \mid \omega_i^h = 1, \xi_i^h = 0, \beta) P(\omega_i^h = 1, \xi_i^h = 0 \mid \theta_h, \beta) + \\
&\quad P(Z_i^h \mid \omega_i^h = 0, \beta) P(\omega_i^h = 0) \\
&= p_i^h s_i^h\, (\theta_h \tau_h)\, \pi_i^h + a_i^h 1/3 (1 - \theta_h \tau_h) \pi_i^h + o_i^h 1/3 (1 - \pi_i^h) \\
&= \underbrace{\pi_i^h \big(\underbrace{\theta_h \tau_h p_i^h s_i^h}_{\text{present}} + \underbrace{(1 - \theta_h \tau_h) a_i^h 1/3}_{\text{absent}} \big)}_{\text{correctly mapped}} + \underbrace{(1 - \pi_i^h) o_i^h 1/3}_{\text{incorrectly mapped}} .
\end{aligned} \tag{38}$$

Analogously, while slightly more involved due to the purity considerations causing $Z_j^t = (R_i^t, S_i^t)$ to depend on both θ_h and θ_c, we compute

$$\begin{aligned}
P(Z_i^t \mid \theta_h, \theta_c, \beta) &= \int_{\xi_j^t, \omega_j^t} P(Z_j^t \mid \xi_j^t, \omega_j^t, \beta) \times P(\xi_j^t, \omega_j^t \mid \theta_h, \theta_c)\, d(\xi_j^t, \omega_j^t) \\
&= P(Z_j^t \mid \omega_j^t = 1, \xi_j^t = 1, \beta) P(\omega_j^t = 1, \xi_j^t = 1 \mid \theta_h, \theta_c) + \\
&\quad P(Z_j^t \mid \omega_j^t = 1, \xi_j^t = 0, \beta) P(\omega_j^t = 1, \xi_j^t = 0 \mid \theta_h, \theta_c) + \\
&\quad P(Z_j^t \mid \omega_j^t = 0, \beta) P(\omega_j^t = 0) \\
&= \pi_i^t \Bigg(\underbrace{\alpha \big(\underbrace{\theta_c \beta \tau_t p_i^t s_i^t}_{\text{present}} + \underbrace{(1 - \theta_c \tau_t) a_i^t 1/3}_{\text{absent}} \big)}_{\text{from cancer cell}} + \\
&\qquad \underbrace{(1 - \alpha) \big(\underbrace{\theta_h \beta \tau_h p_i^t s_i^t}_{\text{present}} + \underbrace{(1 - \theta_h \tau_h) a_i^t 1/3}_{\text{absent}} \big)}_{\text{from healthy cell}} \Bigg) + (1 - \pi_i^t) o_i^t 1/3,
\end{aligned} \tag{39}$$

with $s_i^t = P(S_i^t \mid \omega_i = 1, \xi_i = 1, \beta)$ analogously to the healthy case above. As can be seen, the integral turns into a sum over the three cases $\{\omega_i = 0\}, \{\omega_i = 1, \xi_i = 0\}$, $\{\omega_i = 1, \xi_i = 1\}$, reflecting that Z_i is either (1) incorrectly mapped, (2) correct and not affected by the variant, or (3) correct and affected by the variant. For computing p_i, a_i, and o_i, we need a linear runtime in the length of the considered window, since we are using a banded pair HMM (see section 5.3.1). However, this only needs to be done once for all likelihood computations in the parameter space. We can therefore infer the following central corollary.

Corollary 1. *$L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ can be computed in $O(k + l)$ operations.*

References

T.S. Alioto, I. Buchhalter, S. Derdak, B. Hutter, M.D. Eldridge, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, page 10001, 2015. doi: doi.org/10.1038/ncomms10001. URL `https://www.nature.com/articles/ncomms10001`.

Manuel Allhoff, Alexander Schnhuth, Marcel Martin, Ivan G. Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC bioinformatics*, 14 Suppl 5:S1, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S5-S1.

R.A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. The causes and consquences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, April 2016.

David W. Craig, Sara Nasser, Richard Corbett, Simon K. Chan, Lisa Murray, Christophe Legendre, Waibhav Tembe, Jonathan Adkins, Nancy Kim, Shukmei Wong, Angela Baker, Daniel Enriquez, Stephanie Pond, Erin Pleasance, Andrew J. Mungall, Richard A. Moore, Timothy McDaniel, Yussanne Ma, Steven J. M. Jones, Marco A. Marra, John D. Carpten, and Winnie S. Liang. A somatic reference standard for cancer genome sequencing. *Scientific Reports*, 6:24607, April 2016. ISSN 2045-2322. doi: 10.1038/srep24607. URL `https://www.nature.com/articles/srep24607`.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011. ISSN 1546-1718. doi: 10.1038/ng.806.

R Durbin, S Eddy, A Krogh, and G Mitchison. Biological Sequence Analysis. *Current Topics in Genome Analysis 2008*, 1998. ISSN 0263-6484. doi: 10.1017/CBO9780511790492.

D. Earl, K. Bradnam, J. St.John, A. Darling, D. Lin, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21:2224–2241, 2011.

Bjrn Grüning, Ryan Dale, Andreas Sjdin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0046-7. URL `https://www.nature.com/articles/s41592-018-0046-7`.

R.J. Hause, C.C. Pritchard, J. Shendure, and S.J. Salipante. Classification and characterization of microsatellite instability across 18 cancer types. *Nature Medicine*, 22(11):951–959, 2016. doi: doi.org/10.1038/nm.4191. URL `https://www.nature.com/articles/nm.4191`.

Johannes Köster and Sven Rahmann. Snakemakea scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480. URL `https://academic.oup.com/bioinformatics/article/28/19/2520/290322`.

Samuel Levy, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, Sep 2007. doi: 10.1371/journal.pbio.0050254. URL `http://dx.doi.org/10.1371/journal.pbio.0050254`.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*, March 2013. URL `http://arxiv.org/abs/1303.3997`. arXiv: 1303.3997.

Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, October 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu356. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4271055/`.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324. URL `http://www.ncbi.nlm.nih.gov/pubmed/19451168`.

Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473 –483, 2010. doi: 10.1093/bib/bbq015. URL `http://bib.oxfordjournals.org/content/11/5/473.abstract`.

Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, November 2008a.

Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, November 2008b. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.078212.108. URL `http://genome.cshlp.org/content/18/11/1851`.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

F. Liu, M.J. Bayarriy, and J.O. Bergerz. Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Research*, 18(2):298–309, 2008.

Ion Mandoiu and Alexander Zelikovsky. *Computational Methods for Next Generation Sequencing Data Analysis*. Wiley, 1 edition, September 2016.

Tobias Marschall, Ivan G. Costa, Stefan Canzar, Markus Bauer, Gunnar W. Klau, Alexander Schliep, and Alexander Schönhuth. CLEVER: clique-enumerating variant finder. *Bioinformatics*, 28(22):2875–2882, November 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts566. URL `http://bioinformatics.oxfordjournals.org/content/28/22/2875`.

Tobias Marschall, Iman Hajirasouliha, and Alexander Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, 2013.

Y.E. Maruvka, K.W. Mouw, R. Karlic, P. Parasuraman, A. Kamburov, et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nature Biotechnology*, 35(10):951–959, 2017. doi: 10.1038/nbt.3966. URL `https://www.nature.com/articles/nbt.3966`.

P. Mueller, G. Parmigiani, C. Robert, and J. Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Society*, 99(468):990–1001, 2004.

Gene Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, 46(3):395–415, May 1999. ISSN 0004-5411. doi: 10.1145/316542.316550. URL `http://doi.acm.org/10.1145/316542.316550`.

G. Narzisi. Lancet: Micro-assembly somatic variant caller. See https://github.com/nygenome/lancet.

Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Benjamin Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 2018. doi: 10.1101/201178. URL `https://www.biorxiv.org/content/early/2018/07/24/201178`.

Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, September 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts378. URL `http://bioinformatics.oxfordjournals.org/content/28/18/i333`.

James T. Robinson, Helga Thorvaldsdttir, Aaron M. Wenger, Ahmet Zehir, and Jill P. Mesirov. Variant Review with the Integrative Genomics Viewer. *Cancer Research*, 77(21):e31–e34, November 2017. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-17-0337. URL `http://cancerres.aacrjournals.org/content/77/21/e31`.

Kristoffer Sahlin, Mattias Frånberg, and Lars Arvestad. Correcting bias from stochastic insert size in read pair data — applications to structural variation detection and genome assembly. Technical Report 023929, bioRxiv, Aug 2015.

C.T. Saunders, W. Wong, S. Swarny, J. Becq, L.J. Murray, and K. Cheetham. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

K. Trappe, A.K. Emde, H.C. Ehrlich, and K. Reinert. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics*, 2014.

J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw Mills, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, and The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238–244, January 2016. ISSN 1061-4036. doi: 10.1038/ng.3489. URL `http://www.nature.com/doifinder/10.1038/ng.3489`.

R. Wittler, T. Marschall, A. Schönhuth, and V. Mäkinen. Repeat- and error-aware comparison of deletions. *Bioinformatics*, 31(18):2947–2954, 2015.

A Uniqueness and computation of the maximum likelihood estimate

The likelihood function of θ_h, θ_c, and β given the data $\boldsymbol{Z}^h$ and $\boldsymbol{Z}^t$ as shown in equation (1) is a higher-order polynomial, which makes it infeasible to derive its maximum analytically. We show in this section, however, that under weak conditions (as given in the following theorem) the likelihood function attains a unique global maximum on the unit interval for each value of θ_h and β. We, in addition, show that the loglikelihood function is strictly concave, which simplifies the numerical maximization.

Theorem A.1. *The likelihood function $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^t)$ (where θ_h and β are fixed) attains a unique global maximum $\widehat{\theta}_c$ on the unit interval $\mathcal{U} = [0, 1]$ when*

1. *the likelihood of θ_h, β given the data from the healthy sample must be non-zero, i.e.,*

$$\prod_{i=1}^{k} P(\boldsymbol{Z}_i^h \mid \theta_h, \beta) > 0;$$

2. *the subset*

$$I := \{\theta_c \in \mathcal{U} : P(Z_j^t \mid \theta_h, \theta_c, \beta) > 0 \textit{ for } j = 1, \ldots, l\} \tag{40}$$

is connected and non-empty;

3. *the purity is greater than 0, i.e., $\alpha > 0$ (otherwise the 'tumor' sample would not contain any cancer cells);*

4. *there exists read data Z_j^t for which the alignment probability π_j^t is strictly larger than zero and $p_j \neq a_j$ (i.e., there must exist an observation that with non-zero probability stems from the locus of interest and provided information about the presence or absence of the indel of interest).*

Proof. The likelihood function with θ_h and β fixed can be written in the form

$$L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c) = C \times \prod_{j=1}^{l} P\left(Z_j^t \mid \theta_h, \theta_c, \beta\right) \tag{41}$$

where C is the constant

$$C \equiv \prod_{i=1}^{k} P(Z_i^h \mid \theta_h, \beta).$$

In the case that condition *(1)* is *not* met, $C = 0$. The likelihood $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c)$ equals zero for all θ_c and, therefore, does not attain a unique global maximum.

Suppose condition *(1)* is met ($C > 0$). Let us consider condition *(2)*. Note that $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c)$ = 0 when $\theta_c \notin I$, since for those θ_c's there exists an observation for which the $P(Z_j^t \mid \theta_h, \theta_c\beta) = 0$. The likelihood L is by definition strictly larger than zero when $\theta_c \in I$. Since the function in eq. (41) is an l-th order polynomial and, therefore, continuous, it must attain a global maximum on the interval I.

Suppose condition *(2)* is met. The point $\widehat{\theta}_c$ is a maximum of $L(\theta_h, \cdot, \cdot \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c)$ if and only if it is a maximum of the loglikelihood function

$$\ell\left(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c\right) \equiv \log L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c) = \log C + \sum_{j=1}^{l} \log P(Z_j^t \mid \theta_h, \theta_c, \beta) \tag{42}$$

(with θ_h, β fixed and $\theta_c \in I$) since the logarithm is a monotonic transform. (Note that ℓ is only defined on the subset I). The second order derivative of the loglikelihood with respect to θ_c is found to be

$$\frac{\partial^2 \ell}{\partial \theta_c^2} = -\sum_{j=1}^{l} \left[\frac{\partial P(Z_j^t \mid \theta_h, \theta_c, \beta)/\partial \theta_c}{P(Z_j^t \mid \theta_h, \theta_c, \beta)}\right]^2 \leq 0 \tag{43}$$

indicating that the loglikelihood function is concave. Note that it is strictly concave, i.e., $\partial^2 \ell / \partial \theta_c^2 < 0$, iff there exists an observation z_j^t for which

$$\frac{\partial P(Z_j^t \mid \theta_h, \theta_c, \beta)}{\partial \theta_c} = \alpha \pi_j^t \tau_t \left[p_j^t s_j^t - a_j^t\right] \neq 0. \tag{44}$$

This inequality holds only when $\alpha \neq 0$, $\pi_j^t \neq 0$ and $p_j^t \neq a_j^t$, which constitutes conditions *(3)* and *(4)*.

Suppose I is the non-empty closed set $[a, b]$ on the unit interval. Since the loglikelihood is strictly concave when conditions *(3)* and *(4)* are met, it attains a unique global maximum $\widehat{\theta}_c$ on I. Because the logarithm is a monotonic transformation, $\widehat{\theta}_c$ must be a unique global maximum of the likelihood function as well.

A similar reasoning holds when I is open or half-open. The maximum must lie on the interior of I, since the likelihood function is zero for those endpoints not in I. For example, when I is the open interval (a, b), then $L(\theta_h, a, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c) = L(\theta_h, b, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c) = 0$ while $L(\theta_h, \theta_c, \beta \mid \boldsymbol{Z}^h, \boldsymbol{Z}^c)$ is strictly positive on I. The loglikelihood function is under conditions *(3)* and *(4)* strictly concave on I, therefore, the likelihood function attains a unique global maximum. □

B Supplementary Figures

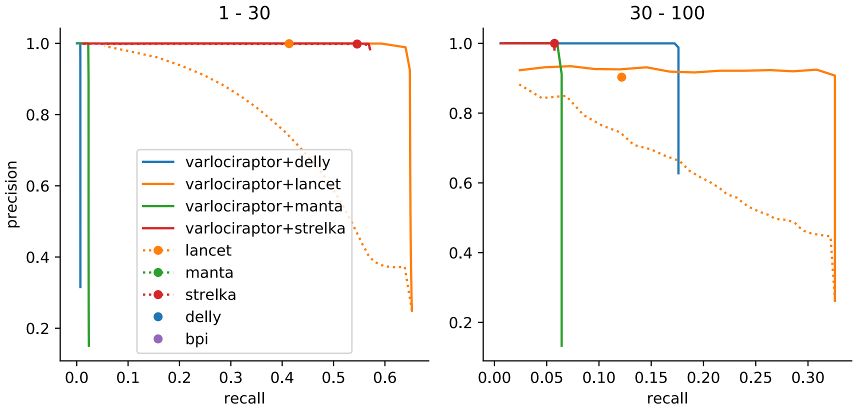


Figure S1: Recall and precision for calling somatic insertions. Results are grouped by deletion length, denoted as interval at the top of the plot. For our approach (Varlociraptor+*) curves are plotted by scanner over the posterior probability for having a somatic variant. For other callers that provide a score to scan over (p-value for Lancet) we plot a dotted line. Ad-hoc results are shown as single dots. Results are shown if the prediction of the caller did provide at least 10 calls. The sharp curves for our approach reflect the favorable property of having a strong separation between the probabilities of true and false positives, see Figure 4.

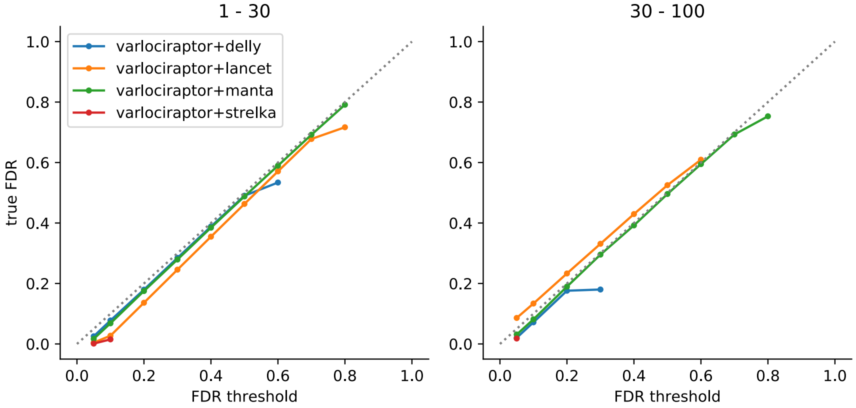


Figure S2: FDR control for somatic insertions. Results are grouped by deletion length, denoted as interval at the top of the plot. The axes denote the desired FDR, provided by the user as input (x-axis), and the true achieved FDR (y-axis). A perfect FDR control would keep the curve exactly on the dashed diagonal. Below the diagonal, the control is conservative. Above the diagonal, the FDR would be underestimated. Importantly, points below the diagonal mean that the true FDR is smaller than the threshold provided, which means that FDR control is still established; in this sense, points below the diagonal are preferable over points above the diagonal.

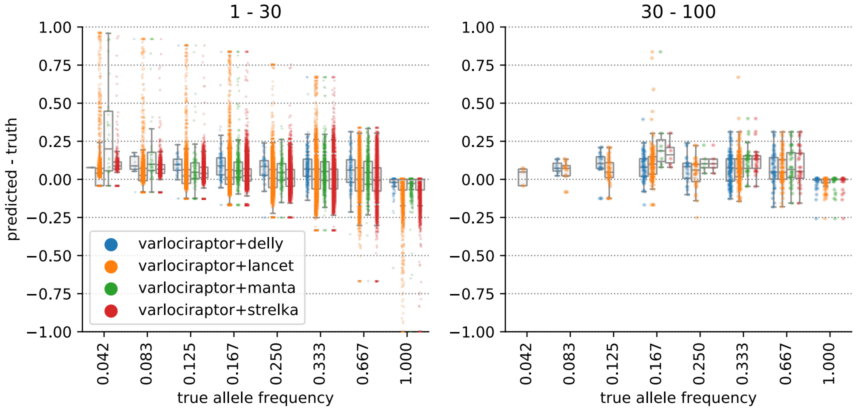


Figure S3: Allele frequency estimation for somatic insertions. Results are grouped by deletion length, denoted as interval at the top of the plot. The horizontal axis shows the true allele frequency, the vertical axis shows the error between predicted allele frequency and truth.

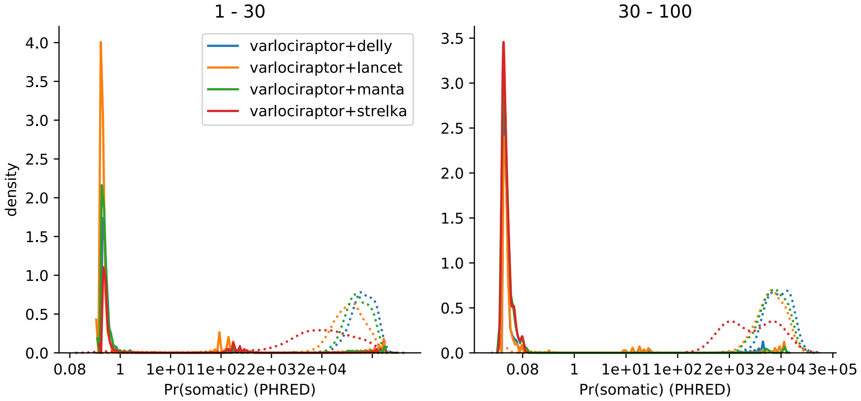


Figure S4: Posterior probability distributions for somatic insertions. Results are grouped by deletion length, denoted as interval at the top of the plot. The x-axis indicates the (PHRED-scaled) probability, and the y-axis indicates relative amounts of calls with this probability. The distributions of posteriors for true positive calls are shown as solid lines, the distributions of posteriors for false positive calls are shown as dotted lines.

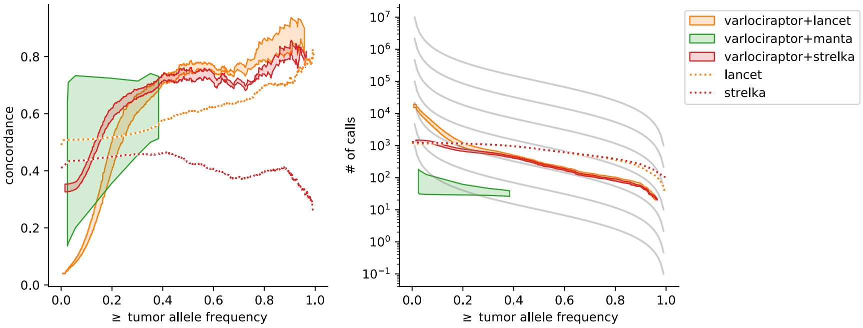


Figure S5: Concordance of somatic insertions on real data. For Varlociraptor, the interval between all calls with a posterior probability of at least 0.9 and at least 0.99 is shown as shaded area. Left: Concordance vs. minimum allele frequency. Right: Number of calls vs. minimum allele frequency. Grey lines depict the theoretical expectation according to Williams et al. [Williams et al., 2016]

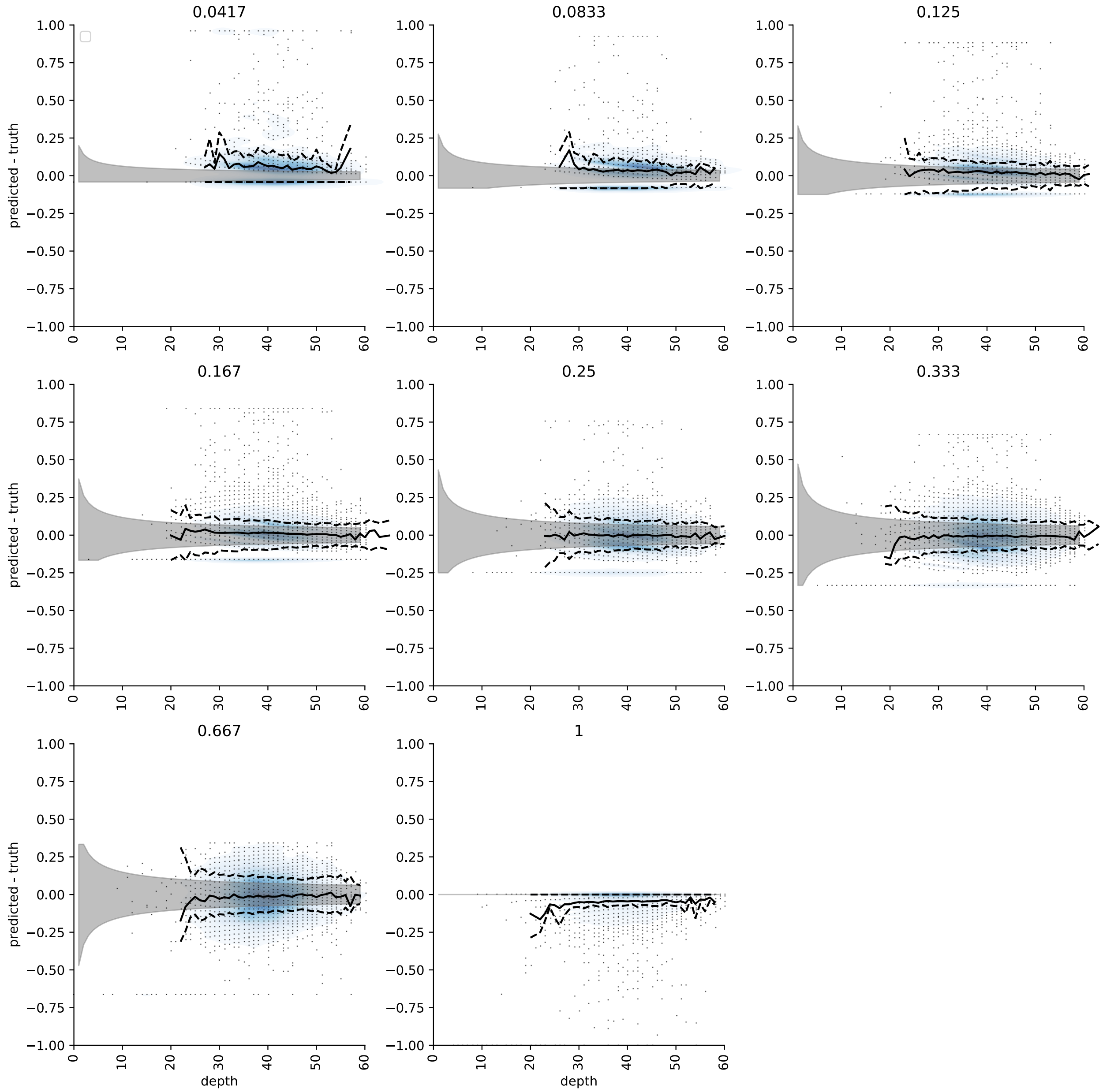


Figure S6: Allele frequency estimation error for somatic deletions compared to sequencing depth. Each plot shows the error (predicted - truth) for a particular true allele frequency (shown above the plot). Dots represent individual predictions, the blue shading shows a corresponding density estimate. The black line shows the mean, the dashed lines depict the standard deviation. The grey area represents the theoretically expected sampling error in an experiment with no further artifacts or biases (the theoretical optimum).

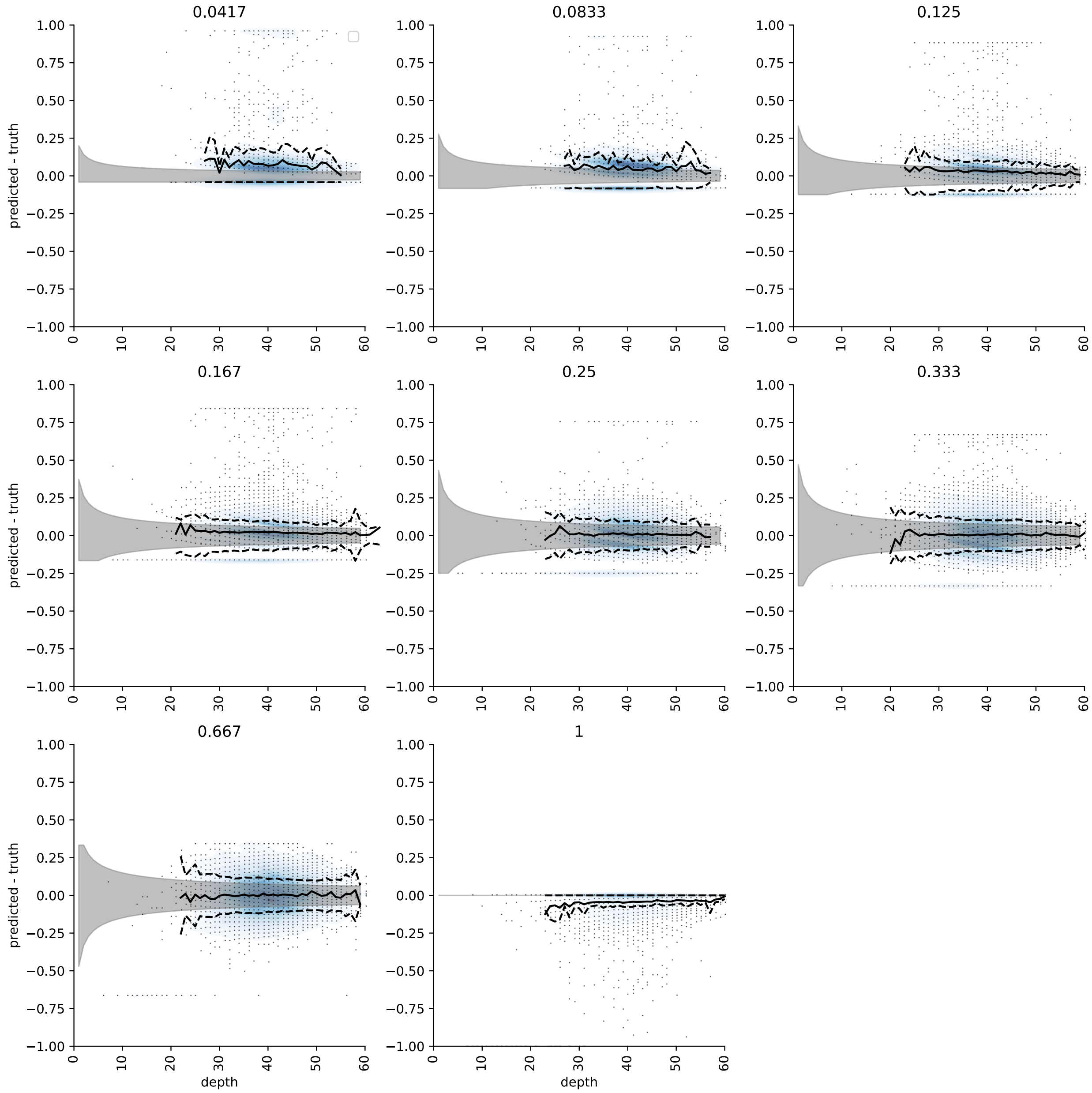


Figure S7: Allele frequency estimation error for somatic insertions compared to sequencing depth. Each plot shows the error (predicted - truth) for a particular true allele frequency (shown above the plot). Dots represent individual predictions, the blue shading shows a corresponding density estimate. The black line shows the mean, the dashed lines depict the standard deviation. The grey area represents the theoretically expected sampling error in an experiment with no further artifacts or biases (the theoretical optimum).

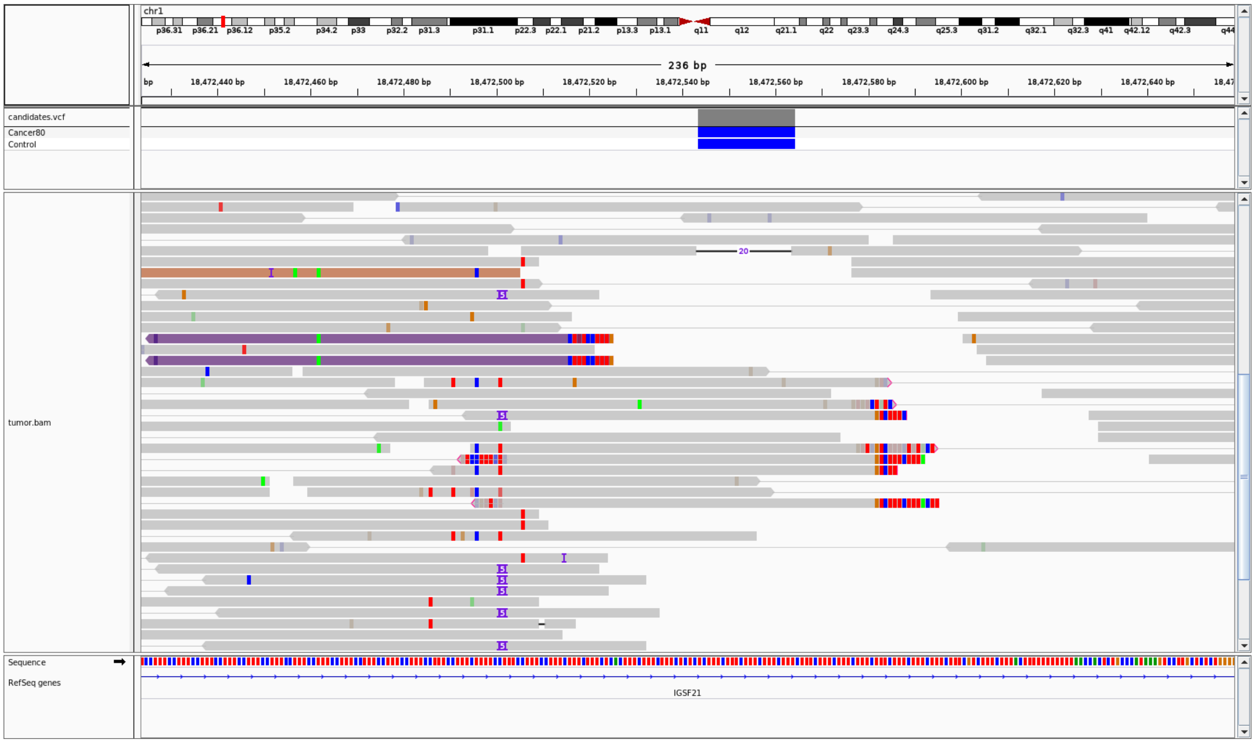


Figure S8: Example of a variant in a repetetive region that causes misplaced softclips, highlighting the need for a realignment against the variant allele (taken from our simulated dataset (see section 2.4). The clipped alignments (shown as mismatches at the read ends) should instead have a 20 bp deletion as the read at the top. Visualization was performed with IGV [Robinson et al., 2017].