2.3 Cancer Diagnostics and Therapy from Molecular Data

Sven Rahmann
Alexander Schramm
Johannes Köster

Abstract: The past decade has seen unprecedented progress in the survival chances of cancer patients as a consequence of new treatments targeting tumor-specific cellular processes, which have been uncovered by molecular genetic analyses. From a data analysis perspective, the main challenge is the high dimensionality and multimodality of the genetic data relative to the small sample sizes (numbers of patients). From a computational perspective, the analysis of high volumes of data (about 100 GB of sequencing data for an individual tumor genome) currently requires high-powered computational resources and still remains challenging in the very short time frames that are desired to start treatment immediately.

We discuss two avenues of progress. First, we present methods that are able to extract most of the genetic variants from a sequenced tumor genome, but require only 2 % to 5 % of the computational resources compared with the current state-of-the-art procedures.

Second, we discuss a versatile unified statistical model for distinguishing true variants from technical artifacts of the DNA sequencing process.

Using analyses of paired samples from primary and relapse neuroblastoma tumors, we are able to extract patterns of tumor evolution that are correlated with cancer progression and the escape of tumors from therapeutic intervention.

As a result, a novel risk classification of neuroblastoma has been established based on genomic and mutational data.

2.3.1 Introduction

Cancer patients nowadays receive precise diagnosis and personalized therapy based on their individual molecular genetic data.

Here, we report on the analysis of DNA data from patients with neuroblastoma, a solid tumor typically occurring in children.

Diagnostics and prognoses are based on DNA sequencing, currently ranging from a few hundred targeted genes to entire genomes requiring 100 GB per patient in the near future.

Identifying relevant variants in the DNA that serve as biomarkers to distinguish between different risk classes, or primary tumors from relapses, or treatable versus non-treatable tumors, is and remains challenging, but every step of progress in this

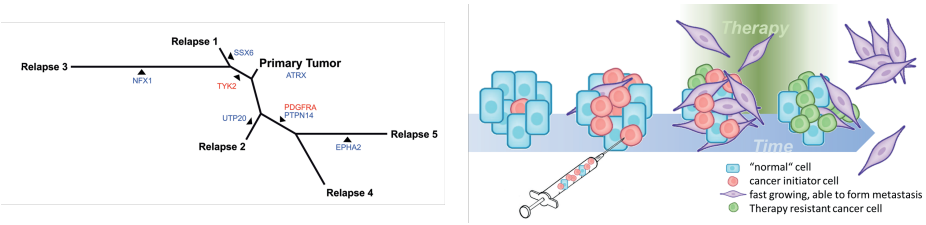

Fig. 2.15: Left: Evolutionary tree derived from binary features (presence/absence of informative single nucleotide variants) of a primary neuroblastoma tumor and several relapse samples taken from different tissues and at different times from the same patient [583]. Right: Illustration of evolution of tumor heterogeneity under therapy over time [588].

field helps to make better treatment decisions in the long run. The following molecular features derived from genomic data are of primary interest: (1) single nucleotide mutations (or variants, SNVs), (2) short insertions or deletions of DNA, (3) large structural variants (e.g., chromosomal translocations), (4) copy number changes (gain or loss of genetic material in tumor cells), (5) epigenetic changes, such as DNA methylation changes, and (6) differences in gene expression.

Over the past years, we have developed feature extraction workflows and data analysis processes for each type of feature mentioned above. For data analysis workflows in general, but especially in medicine, reproducibility of derived data from raw data is of utmost importance. The basis of each of these processes is our workflow management system called Snakemake [345, 450], which is now widely used worldwide, as it guarantees reproducibility in particular for large-scale DNA sequence analysis workflows. Furthermore, the Bioconda package repository [242] was founded by one of us (JK) and now, with widespread community support, acts as a central repository for semantically versioned bioinformatics software, which is made available in a reproducible way.

In the following, for simplicity, we focus on the first type of features (SNVs), but these findings also translate to the other variant types, if appropriately adjusted. In particular, we discuss whether we can determine genomic variants that distinguish primary neuroblastomas from those that re-occur after therapy (referred to as relapses or relapse samples). The latter are responsible for adverse disease courses and are currently considered to be incurable. It was therefore highly encouraging that we were able to identify several genes with recurrent mutations present only in relapse samples [583]. Figure 2.15 summarizes some of our key findings on tumor heterogeneity after relapse (left side) and illustrates the tumor evolution process (right side). It is mainly this developing molecular heterogeneity of tumor cells under treatment that currently prohibits effective long-term therapies.

The main resource constraints for this setting and similar situations are a limited number n of samples (patients) versus an extremely high number p of potential features (e.g., each potential variant in the genome observed in at least one sample).
So we face two challenges in particular:

1. resource-efficient detection of candidates of variants (Section 2.3.2)
2. accurate classification of candidates in each sample (true variant vs. noise, technical artefact, etc.; Section 2.3.3)

2.3.2 Resource-Efficient Detection of Variant Candidates

Standard genetic mutation or variant analysis starts with an extremely compute-intensive step: the localization of every single sequenced DNA fragment (or "read"; there are literally millions of DNA reads in a single dataset) in the genome, and a pairwise comparison between the fragment and the genomic sequence. Such pairwise alignments are the basis of variant calling: many reads showing a certain difference at the same position compared with the reference genome, this provides convincing evidence that the sequenced genome contains a specific genetic variant at that position, either in both inherited chromosome copies (homozygous variant) or in just one (heterogzygous variant). To be precise, complex statistical models and tests are necessary to distinguish true variants from possible technical artifacts (see Section 2.3.3).

This first localization and comparison step is performed by so-called *read mappers*, such as BWA-mem [391], bowtie2 [371], minimap2 [390], or PEANUT [344]. Extensive parallelism on both multi-core systems and GPUs keep the (wall clock) time of this step within a few hours. However, the overall CPU work consists of many CPU days or months for a single dataset, consuming considerable energy.

It is therefore of high interest to develop more resource-frugal methods to achieve the same task, or at least a large fraction of it. We explored *alignment-free* methods as an alternative to the above mapping and alignment-based method. In particular, we propose to use short DNA strings of length k (so-called k-mers) to directly detect potential single-nucleotide variants, as we now describe.

2.3.2.1 Genome Preprocessing

We first preprocess the reference genome.

1. Select an appropriate value for k, such that most k-mers are unique in the reference genome. Our studies indicate that $21 \leq k \leq 31$ works well for the human genome [517].
2. Build a (very large) hash table of k-mers in the human reference genome and the number of times that they occur. We need to take into account that double-stranded DNA is equivalent to its reverse complement.
3. Mark the unique k-mers; they point to a unique position in the genome.
4. Among the unique k-mers, mark those that are *robustly unique* against single substitutions, i.e., those for which no Hamming-distance-1 neighbor also occurs in the genome. The resulting robustly unique k-mers do not only point to a unique

location in the genome; they also cannot easily be changed into *k*-mers that occur at a different genomic location.

The alignment-free methods work either exclusively on the robustly unique *k*-mers or on all unique *k*-mers, giving the information from the robustly unique *k*-mers a higher weight. This pre-processing step has to be performed only once for any genome version.

We found that multi-way bucketed Cuckoo hash tables are ideally suited for the task, as they allow relatively quick construction times and yield very fast lookup times later. There are smooth trade-off options between lower memory usage and even faster lookup times.

In a preliminary study [756], we designed and implemented these hash tables for a simpler application than variant calling: xenograft sorting. Here, a human tumor is engrafted into another organism (typically a mouse) to be able to study its evolution and response to different therapies. When such a tumor is sequenced, one obtains a mixture between human and mouse DNA reads, so that all reads have to be assigned to the organism of origin before proceeding further. This assignment is called xenograft sorting. Even though human and mouse are quite similar on a genetic level, they can be sufficiently well distinguished on the *k*-mer level. We presented a classification method based on *k*-mer hash tables, as outlined above, with extremely high accuracy, but using much less CPU work than previous methods: less than 25 % of comparable hash-based methods and less than 5 % of classical alignment-based methods [756]. We additionally showed that the placement of keys in the hash table can be optimized to yield optimal average look-up times (based on the number of random memory accesses, i.e., likely cache misses), saving 10 % to 15 % of CPU work for each sample (after a 48 CPU hour optimization procedure that has to be run only once) [757].

2.3.2.2 Basic Alignment-Free Variant Calling

The underlying idea of this method is as follows: We count all the *k*-mers in a sequenced sample and produce a histogram of the count values. A typical (unique) *k*-mer should have a copy number of two (in a diploid genome) when no variant is present. We therefore analyze the histogram of observed *k*-mer counts (Figure 2.16) from the sample. The leftmost peak (counts near zero) can be explained by rare *k*-mers due to sequencing errors or contamination; we can attempt to correct these, or ignore them entirely. We fit a negative-binomial mixture model to the remaining peaks occurring at equidistant counts. The main peak corresponds to a copy number of 2 in a diploid genome (from *k*-mers present in both the maternal and paternal chromosome set).

The initial analysis is restricted to the robustly unique *k*-mers from the reference genome. We expect that each such *k*-mer has a copy number of either 0 (homozygous variant), 1 (heterozygous variant) or 2 (no variant) in the sample. Higher copy numbers could be explained by segmental duplications, which we do not consider at this point. If we suspect a variant, we look for *isolated single nucleotide variants*, i.e., *k*-mers with

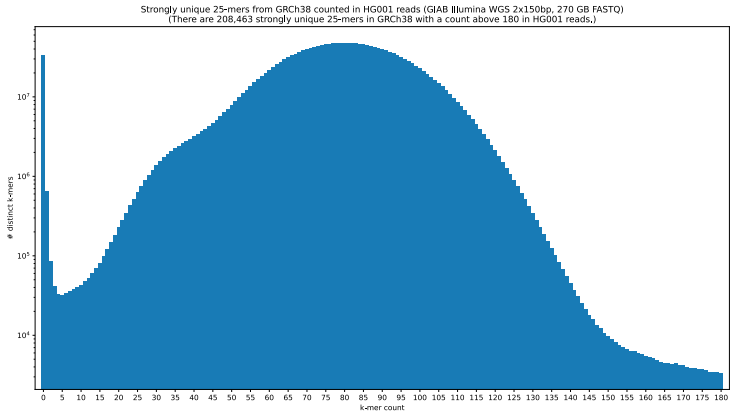

Fig. 2.16: Illustration of a *k*-mer count histogram, relating observed *k*-mer counts (x-axis) to their frequency (y-axis; logarithmic). The leftmost peak (close to zero) represents noise and erroneous *k*-mers, mostly due to sequencing errors. The main peak (near 80, approximately the sequencing coverage of this example) corresponds to the standard copy number of 2. The shoulder (near 40) then corresponds to a copy number of 1 and consists of *k*-mers that are part of heterozygous isolated mutations. This histogram was created from a control sample; in a tumor sample, more irregularities, especially additional peaks at higher copy numbers, can be expected.

a Hamming distance of 1 to the reference *k*-mer, among the *k*-mers in the sample. If we find a unique one (with the expected copy number), we store the pair of reference *k*-mer and modified *k*-mer as a candidate for a variant.

This process can be implemented very efficiently, and in addition, it can be trivially parallelized. It yields candidates for Single Nucleotide Variants (SNVs) that then can be checked by statistical methods (see next section). It can also reliably detect copy number variants on long segments. However, it cannot easily detect more complex variants, such as two SNVs in close proximity, short indels, or structural variants: Here translating *k*-mer information into an exact variant is more difficult, but can resort to alignment-based methods for the local regions around areas with suspicious *k*-mer frequency structure.

Perspectives Alignment-free variant calling is still an active research area, and while we made contributions to the underlying data structures (engineered Cuckoo hash tables) and were successful in calling selected SNVs, further ideas are necessary to call larger classes of variants reliably. Possible approaches include using locality sensitive hashing, in particular min-hashing, instead of exact *k*-mer hashing, combined with *hybrid methods* between alignment-free and alignment-based approaches. To assess the potential of min-hashing-based methods, we conducted a detailed statistical feasibility study, examining when it is useful to include known variants into a *k*-mer-based read mapper (and when not; see [515]), paving the way for novel approaches.

2.3.3 A Unified Statistical Model for Genomic Variant vs. Artifact Classification

We present an extension and generalization of a latent variable model originally published by Köster, Dijkstra, Marschall, and Schönhuth [342].

For this, we consider a set S of samples. Samples can be related with each other in three ways.

1. There can be clonal inheritance between two samples $s_1, s_2 \in S$: sample s_1 inherits all constitutive genetic variants of sample s_2. In addition, the tissues of origin of both samples may have developed their own *somatic* mutations during their lifetime until sequencing.
2. There can be Mendelian inheritance [443] between samples $s_1, s_2, s_3 \in S$: the individual of origin of sample s_1 inherits constitutive genetic variants of two parental individuals (s_2 and s_3).
3. A sample $s \in S$ can be contaminated with a fraction of another sample $s' \in S$.

We represent the three relationships in a directed graph $G = (S, I_c, I_m, C)$ (the *sample graph*) with edge types $I_x \subseteq S \times S$ for clonal ($x = c$) and Mendelian ($x = m$) inheritance as well as $C \subseteq S \times S$ for contamination. The corresponding contamination fraction can be obtained with $c : C \to [0, 1]$.

The above representation can be used to model the three classical cases of genomic variant calling: single-sample or population calling (the graph has no edges) [171], pedigree based family variant calling (Mendelian inheritance edges) [171], and calling of tumor/normal sample combinations (clonal inheritance and contamination edges) [342]. Importantly though, instead of being limited to these, it can reach beyond them by combining the mechanisms into more complex scenarios.

2.3.3.1 Variables and Notation

Observed Variables For each potential genomic variant of interest, we observe sequencing read data $\boldsymbol{Z}_s = (Z_1^s, \ldots, Z_k^s)$. If the read data consists of so-called paired-end reads (each investigated DNA fragment is sequenced from both ends), each observation in $Z_i^s \in \boldsymbol{Z}_s$ is a tuple $Z_i^s \in (\{\text{A,C,G,T}\}^+, \{\text{A,C,G,T}\}^+, \mathbb{N})$, with the first and the second element denoting the nucleotide sequence of the read and the last element denoting the so-called observed insert size, that is, the number of bases from the leftmost to the rightmost covered base when aligning the read pair to the most likely position of origin on the reference genome of the investigated species. If the read data consists of so-called single-end reads (each investigated DNA fragment is sequenced just from one end), each observation $Z_i^s \in \boldsymbol{Z}$ is simply the nucleotide sequence of the read, in other words $Z_i^s \in \{\text{A,C,G,T}\}^+$.

Latent Variables The central readout of our model is the allele frequency in each sample s, denoted as latent variable $\theta_s \in [0, 1]$. For each read observation i, there is

a binary latent variable ξ_i^s with $\xi_i^s = 1$ denoting that the observation originates from the variant allele (i.e., from a genome copy hosting the variant under consideration) and $\xi_i^s = 0$ denoting that the read originates from the reference allele (i.e., from a genome copy hosting exactly the same sequence as in the reference genome of the corresponding species). In addition, a binary latent variable ω_i^s, denoting whether the observation has been aligned to the correct ($\omega_i^s = 1$) location of origin in the reference genome, is used.

Extensions for Bias Estimation The model can be further extended in order to estimate biases in additionally observed properties of the read data, that is, the strand, the read position supporting the variant, the read orientation, and whether the alignment against the reference genome covers the entire read. Biases from an equal distribution in the observed values of variant supporting reads for any of these properties typically indicate an artifact. For clarity and brevity, we omit the integration of these biases in our model here. An integration of strand bias can be already found in [342].

2.3.3.2 Latent Variable Model

In the following, we briefly introduce the latent variable model used for calculating allele frequency likelihoods that has been published recently [342], and then provide a generalization of the method. When evaluating if a read deviates from the reference genome, two types of uncertainty are to be considered. First, there is *alignment uncertainty*: often, a read can be aligned at multiple loci in the reference genome (also see Section 2.3.2).

Depending on their similarity, there is more or less certainty about the optimal positioning of the read. Read mappers and alignment tools, such as BWA [391], report this uncertainty as mapping quality (MAPQ), which can be translated into a probability π_i^s associated with each read observation Z_i^s to be aligned to the correct locus. Second, there is *typing uncertainty*: the observed read sequence is not a perfect representation of the true DNA fragment that has been sequenced, but instead a measurement entailing potential errors and artifacts. The DNA sequencing machine provides an estimate of the certainty of each reported base as the so-called base quality, which can again be translated into a probability of the reported base to be correct. In addition, depending on the sequencing technology, there are known rates of false insertions or deletions of bases in the reported read sequences, as remarked on for example by Schirmer, D'Amore, Ijaz, Hall, and Quince [578].

We now model the relationships between our observed and latent variables, while taking above mentioned uncertainties into account. For each observation Z_i^s in sample s, we handle alignment uncertainty by defining the distribution of the latent variable ω_i as

$$\omega_i^s \sim \text{Bernoulli}(\pi_i^s). \tag{2.3}$$

The distribution of the latent variable ξ_i^s depends on the expected fraction of observations from the variant allele. If s is not contaminated by another sample, we define

$$\xi_i^s \sim \text{Bernoulli}(\theta_s \tau). \tag{2.4}$$

Thereby, $\tau \in [0, 1]$ denotes a sampling bias that occurs because it is usually harder to obtain observations from the variant allele: it is harder to align, and depending on the size of the variant, harder to obtain reads that sufficiently cover it [342]. If, in contrast, s is contaminated by a s' (i.e., $e = (s, s') \in C$) with fraction $\alpha = c(e)$ we define

$$\xi_i^s \sim \text{Bernoulli}(\alpha\theta_s \tau + (1 - \alpha)\theta_{s'} \tau). \tag{2.5}$$

In other words, the expected fraction of observations from the variant allele becomes a mixture of the allele frequencies in s and s'.

Then, typing uncertainty can be modeled as

$$Z_i^s \mid \xi_i^s, \omega_i^s \sim \begin{cases} p_i \text{ if } \xi_i^s = 1, \omega_i^s = 1 \\ a_i \text{ if } \xi_i^s = 0, \omega_i^s = 1 \\ o_i \text{ if } \xi_i^s = 0, \omega_i^s = 0. \end{cases} \tag{2.6}$$

Here, a_i, p_i, and o_i are probability distributions modeling the case that the observation comes from a genome copy where the variant is present (p_i), absent (a_i), or from a different locus (o_i). These can be computed using Pair Hidden Markov models, which essentially realign the read sequence against the sequence of reference and alternative allele while statistically considering sequencing error rates, as shown in Köster, Dijkstra, Marschall, and Schönhuth [342] for deletions and insertions. Since then, via analogous approaches, our model has been extended to also support all other common variant types ranging from small (SNV, MNV) to structural variants such as inversions, duplications, and arbitrary chains of breakpoints.

By combining the above relations, the model can be used to calculate the likelihood of a given combination of allele frequencies of samples $S = \{s_1, \ldots, s_n\}$ as

$$\Pr(\boldsymbol{Z}_{s_1}, \ldots, \boldsymbol{Z}_{s_n} \mid \theta_{s_1}, \ldots, \theta_{s_n}) = \prod_{j=1}^{n} \prod_{i=1}^{|\boldsymbol{Z}_{s_j}|} \Pr(Z_i^{s_j} \mid \theta_{s_1}, \ldots, \theta_{s_n}) \tag{2.7}$$

while assuming independence between the read observations. Note that the computation of the likelihood function is linear in the total number of read observations, as we have shown previously [342].

2.3.3.3 Prior Distribution

The prior probability of a given allele frequency combination $\theta_{s_1}, \ldots, \theta_{s_n}$ in our generalized model can be computed by considering the dependencies between the samples modeled by the sample graph G (see beginning of Section 2.3.3). In addition, we assume

that for each sample $s \in S$, a ploidy $\rho_s \in \mathbb{N}$ (which may differ by chromosome, e.g., it may be sex-specific), a somatic effective mutation rate $\mu_s \in [0, 1]$, and a germline mutation rate $\nu_s \in [0, 1]$ are known. For calculating a prior probability, the key is to explain the total allele frequency θ_s by a germline allele frequency ι_s and a somatic allele frequency $|\theta_s - \iota_s|$. Usually, one of the two will be zero, such that variants are explained either by germline or by somatic mutation, but combinations thereof can also happen in rare cases. From the known ploidy ρ_s of a sample $s_j \in S$, we can calculate the set of possible germline allele frequencies $\zeta_s \subseteq [0, 1]^{\rho_s+1}$. For example, for ρ_s, we obtain $\zeta_s = \{0, 0.5, 1\}$; in other words, any germline variants may occur either in no, one allele (0.5 or 50 %), or two alleles (1.0 or 100 %). The prior probability can then be calculated by recursively exploring all possible explanations of a given total allele frequency combination.

For a combination of germline and somatic allele frequencies we can then distinguish between the following cases:

1. All samples that are not direct descendants of other samples (have no incoming edges in I_c and I_m in the graph G) are considered a population and the prior probability of their combination of germline allele frequencies is calculated, as defined by DePristo et al. [171], based on a so-called heterozygosity (i.e., the expected proportion of heterozygous sites in the genome), which is usually known for the investigated species.
2. For any sample $s \in S$ that inherits clonally from another sample $s' \in S$, we calculate the prior probability for the somatic allele frequency $f = |\theta_s - \iota_{s'}|$ according to the method of Williams, Werner, Barnes, Graham, and Sottoriva [730], who report a formula for the expected cumulative number of somatic mutations per frequency. The latter can be translated into the corresponding density by normalizing with the genome size g and taking the first derivative, resulting in $h(f) = \frac{\mu}{f^2 \cdot g}$ for $f > 0$. In order to also be able to calculate the probability for $f = 0$, we define a reasonably small ϵ and define $h(0) = 1 - \int_{\epsilon}^{1} h(f)\mathrm{d}f$.
3. For any sample $s \in S$ that inherits in a Mendelian [443] way from two parents $s' \in S$ and $s'' \in S$, we first calculate the number of expected constitutive alternative alleles in the child and the parents by multiplying the ploidy with the respective germline allele frequency, i.e., $\rho_s \cdot \iota_s$. We then sum over the probabilities of all cases of inheriting chromosomes with or without the variant allele from the parent samples that could explain the expected constitutive alternative alleles. The individual probabilities can be calculated by modeling an urn drawing process without replacement, yielding a hypergeometric distribution. Finally, additional somatic variation, i.e., cases where $\theta_{s_j} - \iota_{s_j} \neq 0$, are handled by multiplying the corresponding prior probability for the somatic allele frequency.
4. Finally, sometimes it might not be possible to formulate prior assumptions about allele frequencies of a sample $s \in S$. In such cases we specify an allele frequency universe $U_s \subseteq [0, 1]$ for a sample and assume a uniform distribution.

By taking the product over the priors for individual or groups of samples derived from distinguishing the above three cases, the prior probability for any combination of germline and somatic allele frequencies can be obtained.

2.3.3.4 Variant Calling Grammar

The above model is implemented in the software Varlociraptor (https://varlociraptor.github.io). Varlociraptor offers a *variant calling grammar* that allows to define a scenario that configures all aspects of the model (prior parameters, sample graph) via a textual representation in YAML format (YAML Ain't Markup Language; https://yaml.org/). A scenario consists of the following sections.

Species In this section, general prior knowledge about the investigated species is defined, such as the heterozygosity (see Section 2.3.3.3) and the ploidy (number of chromosome copies in a cell). The latter may be defined with sex-specific exceptions (such as the X and Y chromosome distribution in humans).

Samples In this section, the samples and their dependencies (i.e., the sample graph) are defined. For each sample, it is necessary to either define an allele frequency universe (leading to a uniform prior across the defined frequencies) or the sex. In the latter case, ploidy and heterozygosity are taken from the species definition and used to configure the prior accordingly. Each sample may be annotated with a contamination by another sample in a given fraction (this can be used to define the common case of having a tumor sample that also contains healthy normal tissue). Finally, each sample may define a type of inheritance (Mendelian or clonal), while referring to the corresponding parental samples.

Events The heart of a scenario is formed by the definition of mutational *events* of interest. These can be used to define any kind of Boolean logic expressions over allele frequencies (discrete or intervals) in the given samples.

An example for a scenario modeling the calling of variants in a patient for which a normal healthy blood sample, a tumor sample, and a relapse sample is used can be seen in Figure 2.17. Here, for simplicity, we have initially not defined any prior knowledge regarding mutation rates etc., thereby modeling a uniform distribution over the defined allele frequency universes. An equivalent scenario including this kind of prior knowledge is shown in Figure 2.18. Here, it can be seen that we are able to define inheritance between the normal and the tumor sample. For the relapse sample, although in principle it should inherit mutations from the tumor sample, it is unknown to what extent this happens, because usually only one or a few subclones survive the therapy. Hence, we refrain from specifying an inheritance between the tumor and the relapse, and instead impose a uniform prior on the possible allele frequencies in the relapse sample.

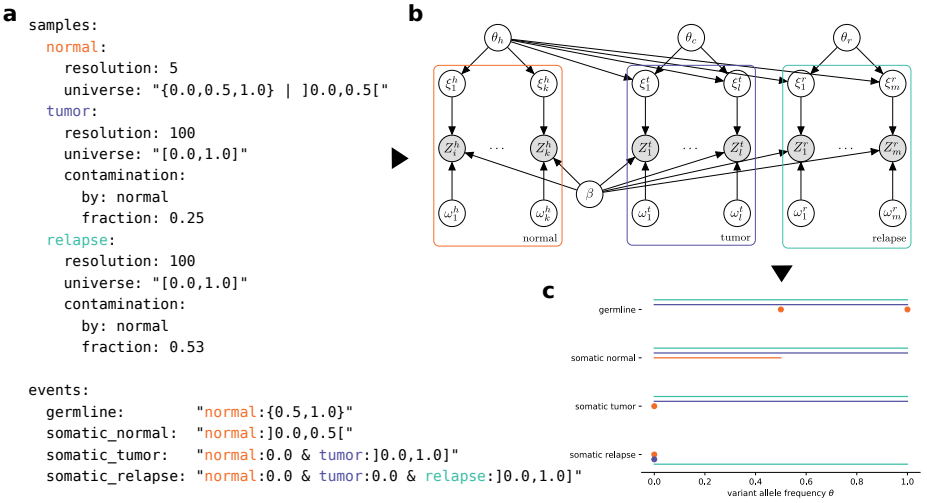

Fig. 2.17: Example of a Varlociraptor scenario specification to distinguish between germline variants and those occurring as somatic events in the primary or relapse sample. (a) Scenario definition via Varlociraptors variant calling grammar. The first section defines the three involved samples normal healthy blood, primary tumor, and relapse after therapy, along with their contaminations and expected allele frequency universe. The second section defines the events of interest via Boolean logic formulas. (b) The resulting structure of the latent variable model, automatically derived from the scenario definition. (c) Visualization of the expected allele frequencies in the three samples for each defined event.

2.3.4 Application and Results

It was previously shown that Varlociraptor is able to significantly improve the recall, while precisely controlling the false discovery rate without the need to tune any technical filter parameters in the absence of a biological interpretation [342]. Here, we illustrate the application of the model by re-analyzing the aforementioned previously published neuroblastoma dataset [583]. In this manuscript, we analyzed genomic data from 17 neuroblastomas, for which DNA was available from the primary tumor and the tumor at relapse. Obtaining the sequence of the entire coding region of the human genome (usually referred to as the "exome") was especially useful for modeling intra-tumor heterogeneity and clonal tumor evolution.

We use the normal-tumor-relapse model formulation from Figure 2.18 and parametrize it as follows. The effective somatic mutation rate in the tumor sample is set to $2.93 \cdot 10^{-6}$. This roughly models the expectation of at most 100 de-novo somatic mutations in typical neuroblastoma tumors found in our original study [583].

Since somatic mutation can also appear in the normal tissue, we set the corresponding effective somatic mutation rate to $2.8 \cdot 10^{-7}$, as reported by Oota [485]. Finally, the

```
species:
  heterozygosity: 0.001
  genome-size: 3.1e9
  ploidy:
    female:
      all: 2
      X: 2
      Y: 0
    male:
      all: 2
      X: 1
      Y: 1

samples:
  normal:
    sex: female
    somatic-effective-mutation-rate: 2.8e-7
  tumor:
    sex: female
    somatic-effective-mutation-rate: 2.93e-6
    inheritance:
      clonal:
        from: normal
    contamination:
      by: normal
      fraction: 0.1
  relapse:
    resolution: 100
    universe: "[0.0,1.0]"
    contamination:
      by: normal
      fraction: 0.53

events:
  germline:        "normal:{0.5,1.0}"
  somatic_normal:  "normal:]0.0,0.5["
  somatic_tumor:   "normal:0.0 & tumor:]0.0,1.0]"
  somatic_relapse: "normal:0.0 & tumor:0.0 & relapse:]0.0,1.0]"
```

Fig. 2.18: Extension of the Varlociraptor scenario specification in Figure 2.17 to include prior knowledge. We define the species (here *Homo sapiens*) in terms of genome size, heterozygosity (expected fraction of heterozygous loci), and sex-specific ploidy (number of chromosome copies). In addition, we model known somatic mutation rates, and define that the tumor inherits germline mutations from the normal sample.

tumor and the relapse sample tissue is usually contaminated by healthy cells. We use the amounts of contamination reported in the original study [583].

Workflow Analyzing sequencing data for genomic variants entails a variety of steps, which we outline in Figure 2.19. The entire analysis is implemented as a Snakemake workflow [343].

First, raw reads are processed by (a) trimming so-called sequencing adapters, (b) mapping them to the reference genome of the corresponding species, (c) removing putative duplicates from the Polymerase Chain Reaction (PCR), and (d) recalibrating base qualities. Sequencing adapters (a) are non-biological artifacts of the sequencing process. Since they are known beforehand, they can be removed from the reads by

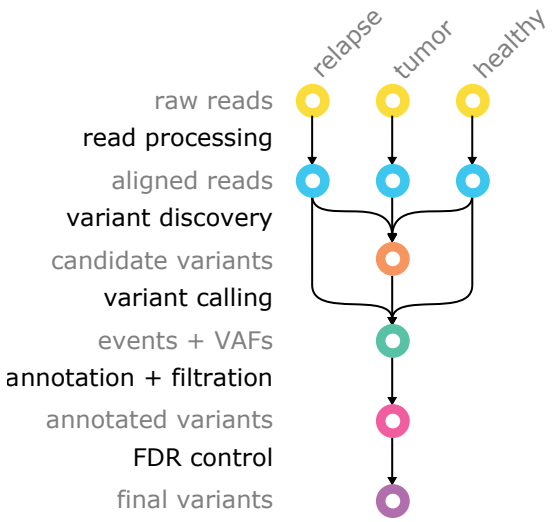

Fig. 2.19: Schematic representation of applied genomic variant calling workflow. Nodes represent original or derived data (gray labels in the left column), arrows represent processing steps (black labels in the left column).

performing an error-tolerant alignment between each read and the known sequence. We use Cutadapt to perform this step [431]. By mapping reads to the reference genome (b), we obtain the correct order and individual differences of each read compared with the representative genome of the underlying species. The resulting read alignments already contain all necessary observations for applying the Varlociraptor model. In order to obtain a signal of sufficient strength, sequencing protocols often entail the amplification of the DNA material via polymerase chain reaction [24]. The result is that there can be multiple reads from the same DNA fragment. Since Varlociraptor assumes each read to be an independent observation, it is important to remove such putative PCR duplicates, which we achieved using Picard tools [500]. Finally, the sequencing process sometimes causes artifacts to appear next to certain motifs [19]. In (d), we therefore use the base recalibration process from the Genome Analysis ToolKit (GATK [171]), which systematically investigates base alteration causing motifs and recalibrates the per base confidence scores in each sequencing read to reflect the uncertainty about whether an altered base is a true signal or a motif-induced artifact.

Second, the aligned reads are used to generate candidate variants. We use the tools Freebayes [222] and Delly [523] for this purpose. While the former covers small variants that can be covered by a single read (SNVs, MNVs, small insertions and deletions), the latter covers large, structural variants (large insertions and deletion, inversion, and duplications). Importantly, while both Freebayes and Delly provide their own statistical models for calling variants, we utilize them to generate candidate hypotheses. Both

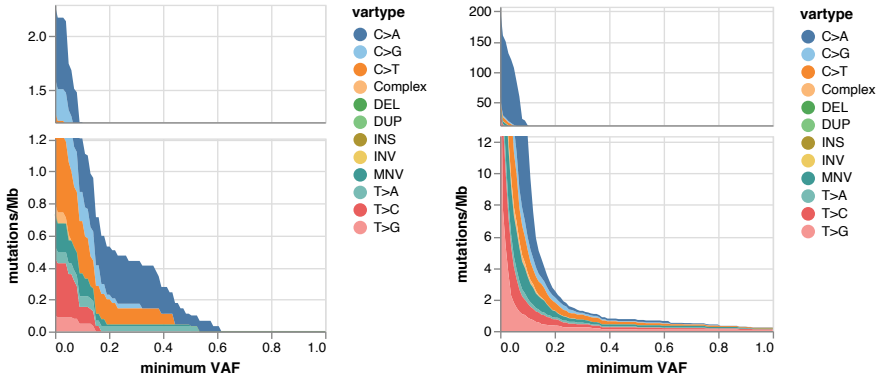

Fig. 2.20: Mutational burden of patient 1 from Schramm et al. [583] in the primary tumor (left) and relapse sample (right). The horizontal axis shows the minimum allele frequency, vertical axis shows the mutational burden as number of coding somatic mutations (calculated as expected value over the posterior probability for having a somatic mutation) per megabase of coding genome. The colors represent different types of mutations (see legend).

models are designed only for specific cases and are not generic enough to handle the composition of samples available in this dataset.

Third, we use Varlociraptor to (a) extract observations for each sample and each candidate variant and (b) apply the model as defined in the corresponding scenario for each patient in the study data.

Fourth, we (a) annotate the variant calls from Varlociraptor with their impact on proteins via the VEP tool [440] and (b) filter them for those that are of interest. In this case, we strive for three disjoint sets of variants

1. Variants that have been previously described as pathogenic or likely pathogenic in other studies.
2. Variants with high impact on the protein but which have not been previously described by other studies.
3. Variants with moderate impact on the protein but which have not been previously described by other studies.

Finally, we separately control the local false discovery rate for somatic variants in either the tumor or the relapse sample on each of the three sets.

Insights In the following, we summarize the most important insights from reanalyzing the study data with this workflow.

Figure 2.20 shows the mutational burden as a curve over the minimum allele frequency on an example patient. It can be seen that the burden for higher frequencies in general increases in the relapse sample compared with the tumor sample. This supports the hypothesis that the relapse sample originates from a subclone of the tumor

sample, which has survived therapy. Thus, one can expect that resistance-inducing mutations in the relapse sample become more abundant. Our findings contribute to the emerging view of resistance to cancer therapies as an evolutionary process. Selection of surviving clones results in mutational fingerprints that are specific for resistant or recurrent tumors. A better understanding of these genetic fingerprints is a prerequisite for identifying markers allowing early detection of resistance or tumor recurrence and enabling timely adjustment of therapies to further improve the survival and cure of cancer patients.

Future work entails the interpretation of individual recurrent deleterious gene and pathway alterations across the analyzed samples. Moreover, we aim to further improve the prior model of Varlociraptor such that assumptions about subclonal inheritance patterns can be incorporated as well.

Finally, we will combine the statistical approach of Varlociraptor with alignment free methods, as outlined in Section 2.3.2. Since Varlociraptor has to perform a realignment of read sequences anyway (see Section 2.3.3.2), we may replace the initial read alignment with an alignment free approach that yields a rough positioning of reads on the reference genome so that they can be selected for validating a given candidate variant with Varlociraptor. For this, it is necessary to accurately estimate the alignment uncertainty from the k-mer hits via, say, the strategy proposed in our previous work on PEANUT [344]. Finally, the detection of candidate variants with alignment free methods has to be extended beyond single nucleotide variants. Here, a possible strategy might be a hybrid approach where aberrations in k-mer counts are translated into an exact variant call by (a) collecting the causing reads, (b) assembling them into one or more consensus sequences [114], and (c) aligning these against the reference genome to determine the nature of the variant.