Internal Report CS Bioinformatics Track 16-01 September 2016

Leiden University

Computer Science

Bioinformatics Track

Reconstructing the subclonal evolution of tumors from targeted sequencing data.

Marleen Nieboer

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Reconstructing the subclonal evolution of tumors from targeted sequencing data

Marleen Nieboer
Master thesis TU Delft & Leiden University
Supervisors: Jeroen de Ridder, Lambert Dorssers, Leendert Looijenga, Erwin Bakker
01-09-2016

Abstract

Motivation: Cancer is driven by the accumulation of somatic mutations. Subclones with partly overlapping mutations may form inside the tumor over time as part of an evolutionary process. It is thought that only a subset of subclones may be involved in acquiring treatment resistance. Thus, identifying the mutations of each subclone can benefit research towards developing more effective anti-cancer therapy. Currently, identifying mutations of subclones is difficult due to technological limitations. Typically, tumor samples are a mixture of multiple subclones. When these samples are sequenced, the measurements are averaged across the subclones that were present in the sample, complicating the reconstruction of subclonal evolution of the tumor. Existing methodology for the reconstruction of subclonal evolution are typically limited to identifying only a small number of subclones in a sample. Alternative sampling techniques such as microdissections aim to reduce the number of subclones in a sample, but require a large number of samples to accurately represent the collection of subclones inside a tumor. However, existing methods cannot easily be applied to a large number of samples as these methods require data from whole genome or whole exome sequencing, making this type of analysis financially impractical. Targeted sequencing is a cheaper alternative to whole genome and whole exome sequencing, but no methods have yet been developed that can reconstruct subclonal evolution from data acquired with targeted sequencing.

Results: We present TargetClone, a method to infer the most likely copy numbers, alleles and frequency of subclones in multiple tumor samples and their subclonal evolution tree from lesser allele frequencies measured with targeted sequencing. We demonstrate that the copy numbers and sample frequency can be inferred with accuracies above 90% with low levels of sequencing noise. Furthermore, we apply the method to simulation data and reconstruct the subclonal evolution trees for two testicular germ cell tumors with resistance to chemotherapy.

Availability: An implementation of TargetClone is available at:
https://github.com/targetclone/TargetClone

1 Introduction

Cancer develops through the accumulation of somatic mutations over time [1,2]. Somatic mutations are unique to the tumor and are not present in cells of healthy tissue. Interestingly, not every cell in a tumor bulk contains the same set of mutations. Instead, tumors are often highly heterogeneous populations of cells which each have unique, but partly overlapping, mutation patterns[3–5]. Every population with the same set of mutations is referred to as a subclone[2,6]. New subclones are formed throughout the development of the tumor as part of the process of clonal evolution. New subclones inherit the genome from their precursor, but may gain additional or lose mutations over time (Fig. 1A)[7]. Some combinations of mutations may be harmful for the subclone, which may result in a decreasing frequency of the subclonal cell population[7,8]. Other combinations may be important in the development of the tumor, or in acquiring resistance to therapy[9,10].

Somatic mutations are typically classified as either somatic variations or copy number variations[6]. Somatic variants are Single Nucleotide Polymorphisms (SNPs) that only occur in the tumor cells, and not in healthy cells. A SNP is a polymorphism in the DNA where the nucleotides on one or both chromosomes of a pair is different from the nucleotide in a reference genome (Fig. 1B). The two nucleotides that vary between chromosome pairs and the reference genome are also called SNP alleles. When the nucleotides on both chromosome pairs are the same, but differ from the reference, the SNP is called homozygous. If the nucleotides differ between the two chromosome pairs, the SNP is called heterozygous. We will refer to SNP alleles as alleles in the rest of this text. Alleles are identified in the genome after aligning the genome to a reference genome. The allele that is the same as the allele in the reference is referred to as the reference allele. The other allele is referred to as the variant allele. When a SNP is not observed on a chromosome pair of a healthy cell, but it is found on a chromosome pair of a tumor cell, we refer to the SNP as a somatic variant (Fig. 1C). In contrast, a copy number variation is a change in the number of chromosomal copies and typically affects more than one nucleotide. Copy number variations are characterized by a gain or loss of a chromosomal region. A copy number variation can be measured at SNP alleles, where an increase or decrease in the number of times an allele is present corresponds to a gain or loss of that allele, respectively (Fig. 1D). These copy number variations are then referred to as allelic copy number variations.

Characterizing the stages of tumor growth at which certain mutations are gained and lost can assist in improving our understanding of how tumors

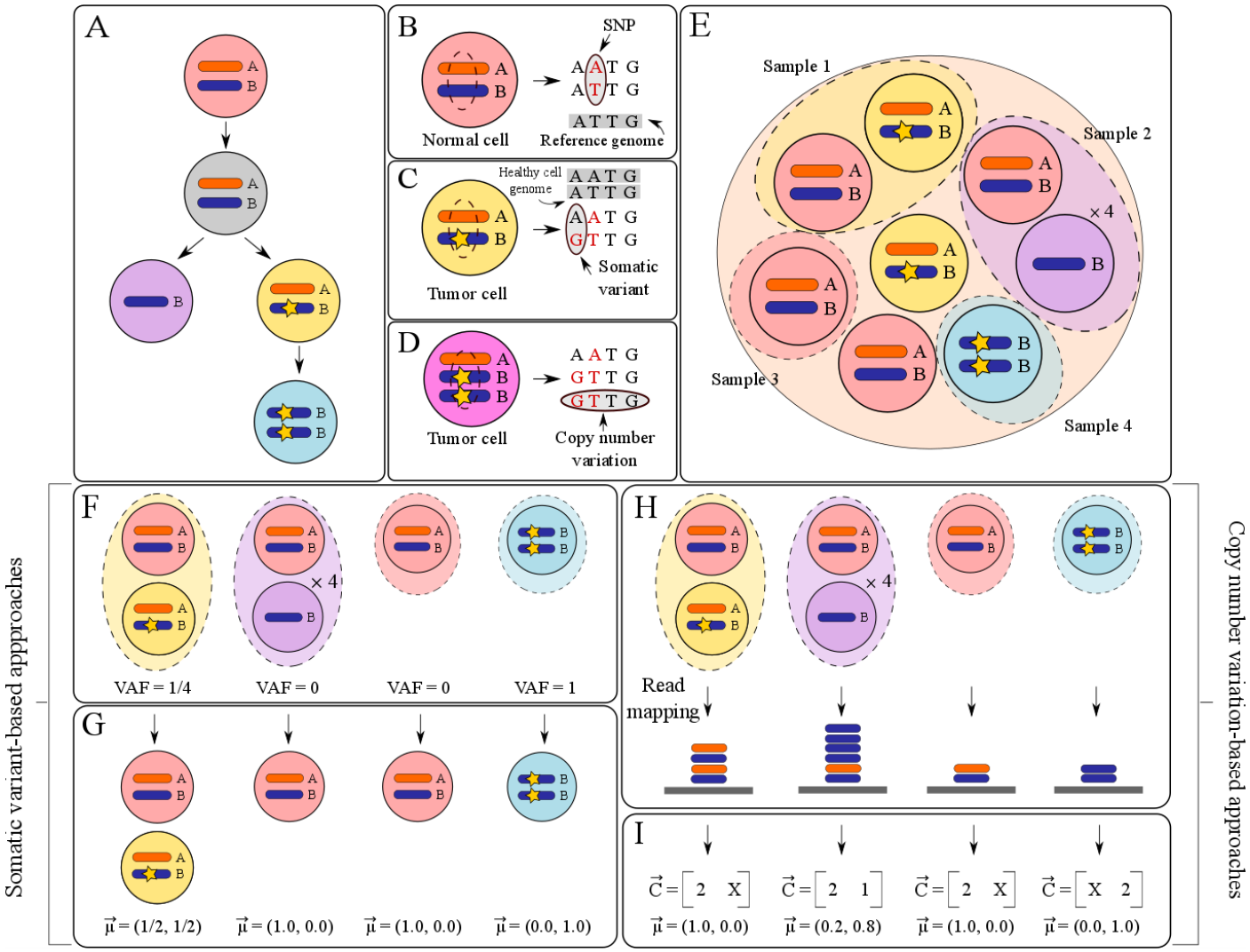

Fig. 1. Illustration of SNPs, somatic variants, copy number variations and how this information is used to characterize tumor heterogeneity. (A) Example of clonal evolution within a tumor. The second subclone (grey circle), has somatic mutations compared to the normal cell precursor in other regions of the genome that are not shown in this figure for clarity. **(B)** Each circle is a cell. The color of a cell indicates that the cell has a unique set of mutations. The red circle is a normal cell, which has two chromosome copies. Each allele is indicated with colored bars, where the orange bar corresponds to the reference allele, and the blue bar corresponds to the variant allele. The reference genome is indicated on the grey bar. We see that a SNP is present in the normal cell. For illustration purposes, we show 3 additional nucleotides around the SNP. **(C)** The grey bars now indicate the alleles of a normal, healthy cell. Compared to the normal cell, the tumor cell has an additional polymorphism from A to G at the first nucleotide, which is a somatic variant. **(D)** In this tumor cell, a copy number variation has occurred and a variant allele has been gained, which is now present in two copies. **(E)** We take four samples (indicated with the dotted lines) from a heterogeneous tumor bulk. At time of sampling, not all subclones necessarily need to be present. In this example, the grey subclone has died out over time. In the purple sample, the purple subclone is actually present 4 times. All the other subclones are sampled once. **(F)** The variant allele frequency (VAF) is shown for every sample. The samples are the same as in Fig. 1E. In the yellow sample, a somatic variant is present on one out of four copies, whereas no somatic variants are present in the purple and red samples. The blue sample has a somatic variant at every copy. **(G)** Expected solutions when we apply somatic variant-based methods to find subclones in the four samples. The arrow points from a sample to the corresponding solution. There is no difference between a normal cell and the purple subclone based on somatic variants, so we do not know the purple subclone is present. **(H)** The sequencing reads of the shown alleles of every sample are mapped to a reference genome, which is shown as the grey bar. Somatic variants are omitted from the mapped reads as this information is not used by copy number variation-based methods. **(I)** Based on the mapped reads, the most likely $\vec{C}$ and $\vec{\mu}$ are reported for each sample. 'X' indicates that the subclone is not present, and thus has no copy number. Again here, as the only difference between the yellow subclone and normal cell is a somatic variant, copy number variation-based methods cannot distinguish between the cells.

develop and may respond to treatment[1,10,11]. However, a couple of limitations in the currently available methodology add challenges to the reconstruction of subclonal evolution.

Technological limitations complicate the reconstruction of subclonal evolution

In the ideal scenario, it would be possible to isolate every subclone and sequence these individually. However, these methods are out of scope of current standard research[9,12,13]. Instead, samples are heterogeneous and typically consist of multiple subclones and normal cell contamination (Fig. 1E). When such mixed samples are sequenced, the measurements will be averaged across the existing subclonal components. As a result, information on the subclonal level is lost. Nevertheless, over the past couple of years a number of methods have been developed to decompose sequencing measurements from these heterogeneous samples into subclones and reconstruct their phylogenetic relations[14].

Previous work

Existing methods for reconstructing (sub)clonal evolution can be categorized into three types: somatic variant-based methods, copy number variation-based methods, and methods that combine both somatic variants and copy number alterations.

Somatic variant-based approaches

Subclones inherit somatic variants from their precursor (Fig. 1A). A commonly made assumption is the infinite sites assumption (ISA)[6,9,15]. This assumption states that a somatic variant will only affect every genomic site once due to the large number of other positions it could have affected. Second, a somatic variant is assumed to never be lost once gained. These assumptions typically restrict the possible number of clonal evolution trees enough to find one or more solutions[15]. However, the assumptions do not hold in the presence of copy number variations. For example, a somatic variant can disappear in a new subclone when the chromosomal region on which it is located is lost. Therefore, most methods restrict their input to only somatic variants that are located in copy number neutral regions. Examples of methods that apply this restriction are TrAp[16] and Clomial[9], which both try to find the most likely decomposition of heterogeneous samples into subclones in the samples based on variant allele frequency (VAF) measurements. The variant allele frequency is a ratio of the variant allele compared to the reference allele, which is measured at positions containing somatic variants (Fig. 1F). TrAp and Clomial use the variant allele frequency measurements to infer a matrix containing a 1 or 0 if a subclone contains a somatic variant or not, respectively. Using the previously named assumptions, the matrix is used to infer the most likely subclonal evolution tree. Rec-BTP makes similar assumptions, but instead tries to find the most likely binary tree given variant allele frequency measurements[17]. PurBayes makes use of Bayesian mixture models to find the most likely clusters of somatic variants in the variant allele frequency measurements, where each cluster represents a subclone[18]. Next to finding subclones, all of these methods also try to infer their frequency in the samples (Fig. 1G). The frequency of *m* subclones in a sample is typically denoted as

$$\vec{\mu} = (\mu_1, \dots, \mu_m) \quad (1)$$

As normal cell contamination is common in tumor samples, the first element of $\vec{\mu}$ will always represent the frequency of the normal cells in the sample. The sum of all elements in $\vec{\mu}$ is 1 by definition.

One common problem that all methods based on solely somatic variants run into is that not all subclones in a sample necessarily need to contain somatic variants. For example, as we can see in Fig. 1F, the purple subclone has no somatic variants and a variant allele frequency of 0, and only contains a copy number variation. As no somatic variants are present in this subclone, the purple subclone is the same as a normal cell for somatic-variant based methods. Therefore, only a normal cell will be inferred with a frequency of 1 in the sample. As a result, copy number variation-based methods have been designed to be applied to the scenarios where somatic variants are not enough to correctly infer subclones.

Copy number variation-based approaches

In certain cancers, the existing somatic variants can mostly be passengers, whereas instead copy number alterations are thought to be the driving force behind tumor growth[2]. Therefore, some methods have been developed which focus on reconstructing subclonal evolution from copy number variations instead. ThetA[19] and TITAN[11] are examples of methods where read depth information from sequencing is used to find the decomposition of a mapped reads into subclones with the highest likelihood. For all samples, these methods map reads from the alleles of subclones in the sample to a reference genome (Fig. 1H). From these reads, the methods try to find the most likely combination of the copy numbers of the subclones and their frequency $\vec{\mu}$ in the samples (Fig. 1I). The copy numbers of the *m* subclones are typically represented in a matrix, here called $\vec{C}$. In the example of Fig. 1I, only one chromosomal region is shown. In a typical study, multiple *(n)* chromosomal regions can be measured. Matrix $\vec{C}$ thus consists of an indication of the most likely chromosomal copy number of subclone *j* at chromosomal region *i:*

$$\vec{C} = \left(C_{i,j}\right) \epsilon\, \mathbb{N}^{n \times m} \quad (2)$$

Similar to $\vec{\mu}$, the first column of $\vec{C}$ will always represent the copy numbers of the normal cells in the sample, which is assumed to be 2. Every other copy number *k* is restricted on the interval [*kmin, kmax]*:

$$\vec{C} = \{C_{*,1} = (2, 2, \dots, 2)^T, \qquad C_{i,j} = \{(kmin, kmin + 1, \dots, kmax)\}$$

From Fig. 1I, we see that the yellow sample will be predicted to only contain normal cells. The reason behind this solution is that since copy number variation-based methods do not include somatic variants in their subclonal reconstruction process, no difference will be observed between the yellow subclone and a normal cell. As both somatic-variant based methods and copy number variation-based methods have the same problem that sometimes no clear difference can be found between subclones if these lack either somatic variants or copy number variations, efforts have been made to combine the two to increase subclonal reconstruction accuracy.

Approaches combining somatic variants and copy numbers

In some cancers, it may initially be unclear what the driving force behind tumor growth is. To overcome this challenge, recently another method called CloneHD[2] has been developed which integrates somatic variants with allele frequencies and copy number information measured across the entire genome and tries to find a decomposition into subclones that is consistent across all data layers using Hidden Markov Models.

Another benefit of combining copy number information with somatic variants is that it can help to overcome the restriction of existing methods to copy number neutral regions. The assumption that all somatic variants are located in copy number neutral regions is common for somatic variant-based methods. The reason why this assumption is made is that it can help resolve ambiguities in how many somatic variants are present. For example, with a variant allele frequency of 0.5, the somatic variant can be present on one out of two copies, but also on two out of four copies. However, knowledge on the number of copies containing a somatic variant is essential in placing a subclone in the correct position in a subclonal evolution tree. It may very well be that a subclone with four chromosome copies is formed later during evolution than a subclone with two chromosome copies, as a subclone with one somatic variant on two copies can potentially be generated directly from a healthy cell. However, if these ambiguities cannot be resolved properly, it is difficult to reconstruct accurate trees of subclonal evolution. Thus, methods typically assume that somatic variants only occur in copy number neutral regions, and in combination with the infinite sites assumption only occur on one of the two chromosome copies. The combination of somatic variants with copy number variations is successfully applied by PhyloWGS[6], PhyloSub[15], PyClone[20], SciClone[5] and EXPANDS[10].

Despite having been reported to produce good results in reconstructing subclonal evolution, most existing methods are restricted to 2 subclones per sample[2], or lose accuracy as the number of subclones increases[6].

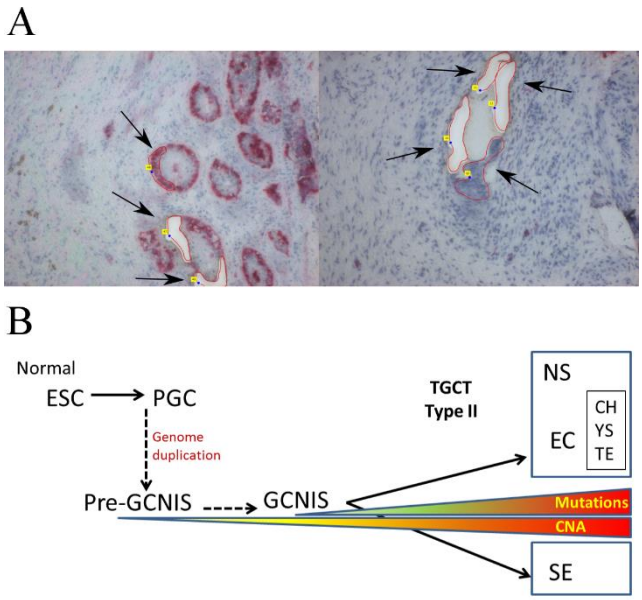

Fig. 2. (A) Example of laser microdissections of a GCNIS component (left) and TE component (right) of a testicular germ cell tumor. The arrows indicate the (circled) regions of interest which will be extracted using a laser. Empty regions of interest indicate that the region has already been extracted. **(B)** Schematic overview of the development of type 2 TGCT. Starting from normal embryonic stem cells (ESC), primordial germ cells (PGC) are formed. When blocked in proper differentiation, the genome of primordial germ cells can be duplicated, forming pre-GCNIS, which further develops into GCNIS after accumulating somatic mutations. GCNIS can develop into seminoma (SE) and non-seminoma (NS). Non-seminomatous tumors are assumed to initially form from a pluripotent embryonal carcinoma (EC), which can further differentiate into chroriocarcinoma (CH), yolk sac tumor (YS) and teratoma (TE). Somatic mutations and copy number variations (CNA) accumulate over time.

Nevertheless, alternative approaches have been developed that aim to overcome the problems introduced by heterogeneous samples by reducing heterogeneity on the level of sampling.

Laser microdissections are useful in obtaining less heterogeneous samples

An alternative method to reduce heterogeneity in samples is through the use of laser microdissections. With this technique, it is possible to stain specific regions of interest in the tumor, which can then be very precisely cut out using a laser (Fig. 2A). The added precision allows for minimization of heterogeneity and normal cell contamination in samples[21,22]. Rather than requiring the deconvolution of heterogeneous tumor samples back into subclones, microdissections can be used to obtain a diverse subset of samples from multiple regions of the tumor in which heterogeneity is aimed to be reduced. However, even with microdissections it has been shown to be difficult to always obtain low levels of normal cell contamination[22]. Therefore, it is still desirable to obtain estimates of the ploidy of the tumor component and the purity of each sample.

Targeted sequencing is a cheaper alternative to WGS and WES

The currently existing methods are not suitable for the task of estimating tumor ploidy and purity for a large number of samples. Estimates of copy numbers are required to apply both the methods that include copy number variations in their model and the methods that require knowledge of which regions are copy number neutral. To obtain high-quality copy number estimates, the methods rely on measurements from either SNP arrays, whole exome sequencing (WES) or whole genome sequencing (WGS)[14]. These types of analysis become financially impractical in studies with a large set of samples, as is typically the case when using microdissections.

An alternative sequencing method is targeted sequencing, which only measures read depth at preselected regions of interest and thereby reduces cost and time of the sequencing process[23]. The idea of being able to reconstruct subclonal tumor evolution from data generated by cheaper sequencing methods is highly interesting. However, cheaper sequencing also comes at the cost of the quality of data that is obtained, which will be explained in the following[24].

Targeted sequencing introduces new challenges

Importantly, it is difficult to obtain accurate copy number estimates from targeted sequencing data. The reason is that measurements are typically distributed in low densities across the genome, meaning that typically not every genomic region is covered. As a result, variation in the read depth due to biases cannot accurately be corrected for through the utilization of the measurements in adjacent regions[25].

The most relevant bias is caused by the varying amounts of PCR cycli required for the typically different amounts of input DNA that are isolated per sample[25]. To correct the variations in read depth resulting from the targeted sequencing process, many (~100) reference samples are required to accurately estimate the variation, which is a costly process. Consequently, targeted sequencing methods usually only result in accurate measurements of the allele frequency at SNP alleles, which is defined as the ratio of reference to variant alleles.

As a result of the loss of correlation between measurements, it becomes difficult to accurately define a phasing of parental alleles. Therefore, it is typical to convert allele frequency measurements to lesser allele frequencies (LAF). The lesser allele frequency is the ratio of the least frequent allele at a SNP position. Let $\overrightarrow{A_c}$ and $\overrightarrow{B_c}$ be matrices of size $n \times m$ containing the number of reference (A) and variant (B) alleles at each genomic region in each subclone in a sample (Fig. 3A). Then the lesser allele frequency in that sample at region i can be computed as (Fig. 3B)

$$LAF_i = \frac{min(\sum \overrightarrow{A_{ci}} \times \vec{\mu},\ \sum \overrightarrow{B_{ci}} \times \vec{\mu})}{\sum \overrightarrow{A_{ci}} \times \vec{\mu} + \sum \overrightarrow{B_{ci}} \times \vec{\mu}} \qquad (3)$$

The lesser allele frequency is usually only measured at positions where the SNP is heterozygous in the normal sample. The reason for this is that it includes the identification of loss of heterozygosity (LOH). LOH is a scenario in which the tumor cell has lost one or more alleles and becomes homozygous, so we measure a lesser allele frequency of 0 (Fig. 3C, third sample), whereas the lesser allele frequency is 1/2 in the normal sample. If the normal cell is also homozygous (so a lesser allele frequency of 0), it would not be possible to observe that the tumor has lost an allele as the lesser allele frequency remains 0 compared to the normal cell.

Some examples of the lesser allele frequency are given in Fig. 3C. The first two samples and the last two samples reveal that it is possible for different combinations of alleles and copy numbers in the tumor cell to result in the same lesser allele frequency. For example, a lesser allele frequency of 1/2 could potentially correspond to having 1 reference and 1 variant allele, but also having 2 reference and 2 variant alleles (Fig. 3C, last two samples). Thus, lesser allele frequencies can correspond to multiple possible copy numbers, which we will refer to as copy number ambiguities. These ambiguities make it difficult to correctly estimate copy number information using lesser allele frequency measurements alone. As a result, it becomes impossible to use existing methods that require copy number information.

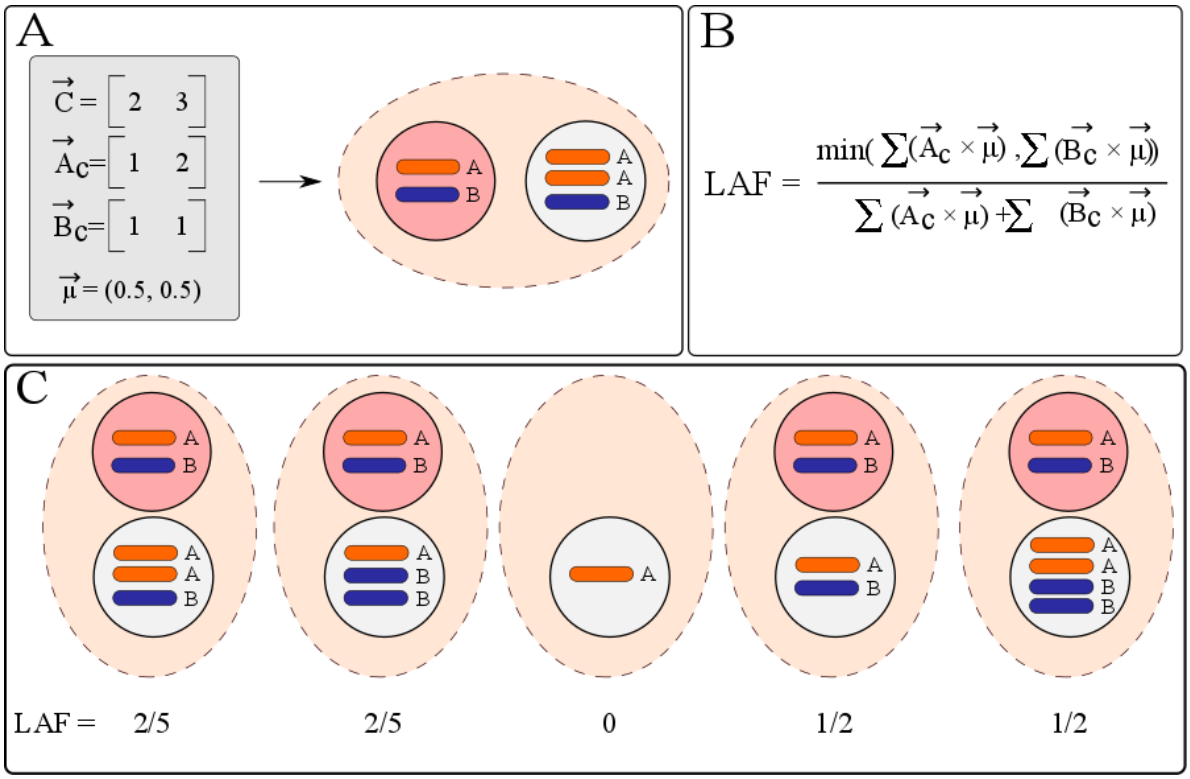

Fig. 3. Computation of the lesser allele frequency (LAF). (A) Example of a sample (dotted line) with two cells, one normal cell with two chromosome copies and one tumor cell with three chromosome copies. The normal cell has one copy of the reference allele A and one copy of the variant allele B, whereas the tumor cell has gained one copy of the reference allele A. **(B)** Formula to compute the lesser allele frequency in a sample using the components shown in Fig. 3A. Formally, the LAF is computed at a chromosomal region *i* (Eq. 3). As this figure focuses on only one region, we omit the region indicator from the figure. **(C)** Example of lesser allele frequency for five samples computed using the formula in Fig. 3B. In each of the samples shown, the tumor cell has a different number of alleles.

Methods that use somatic variants in copy number neutral regions can also not be applied: as a lesser allele frequency of 1/2 can correspond to having 2 or 4 copies and the read depth information is not reliable, there is no way of telling if a somatic variant is located in a copy number neutral region or not from lesser allele frequency measurements alone. However, following the work presented in MEDICC[26], we make use of the assumption of the existence of a minimum number of changes made to the genomes of subclones over time to resolve copy number ambiguities. This is further described in the next section.

Evolutionary relations between samples is assumed to help to resolve copy number ambiguities

All subclones in a tumor are assumed to share evolutionary relations. Mutations are inherited from a precursor and new mutations are gained and lost. We assume that introducing changes into the genome can potentially be harmful for cells as these may occur in essential coding regions[7,8]. Therefore, subclones in which harmful mutations have occurred are expected to show a quick decrease in frequency and are not expected to give rise to new subclones. As such, the existing subclones in a tumor are assumed to not have highly dissimilar genomes, as introducing more mutations increase the chance of the subclonal population dying out. Following this expectation, we assume that the number of somatic mutations that are gained and lost over time is minimal[26]. This assumption is expected to be useful in obtaining better estimates of the copy numbers of subclones as follows. Even if there exist copy number ambiguities at a chromosomal region in one sample, the copy numbers in the other samples may help us resolve such ambiguities. For example, if we measure a lesser allele frequency of 1/2 in sample 1, then that chromosomal region in the tumor subclone could have a copy number of 0, 2 and 4, if we set the maximum allowed copy number to 8. However, we may find substantial evidence that the copy number is 1 for the tumor subclones in the other samples. Therefore, due to the assumption that the number of changes to the genome between subclones is minimal, it is a lot more likely that the tumor subclone in sample 1 also has a copy number of either 0 or 2. However, as a copy number of 0 corresponds to a homozygous deletion where all alleles are lost, which are uncommon as cells can typically not go without most parts of their genome, we can conclude that the most likely copy number of sample 1 in this region is 2. We show in this text what the benefit of the minimum event distance assumption is to reconstruct subclonal evolution from lesser allele frequencies.

Contributions

As explained above, it is very difficult to obtain accurate estimates of copy numbers that would allow the application of existing methods to reconstruct the subclonal evolution in tumors from samples that have been subjected to targeted sequencing. However, as targeted sequencing is cheaper than whole genome or whole exome sequencing, it would be ideal if it were possible to reconstruct subclonal evolution from targeted sequencing data with equal, or higher, accuracy as with whole genome or whole exome sequencing. In this article, we present *TargetClone*, a novel

method which infers the most likely copy numbers and alleles of subclones and their sample frequency from lesser allele frequencies measured in multiple samples using targeted sequencing. Furthermore, a subclonal evolution tree is reconstructed for the inferred subclones. The samples are each assumed to consist of no more than 1 tumor subclone. The inferred ploidy or alleles are optionally combined with measured somatic variants to infer the subclonal evolution tree. Rather than taking as input the measurements of one sample, we demonstrate that our approach of utilizing information across all samples results in a higher accuracy. Furthermore, we apply the method to the case of chemotherapy-resistant type 2 testicular germ cell tumors (TGCT). Type 2 TGCT is a type of germ cell tumors, which are classified into 5 different types[27]. Here, we will focus on only type 2 TGCT.

TGCT is the predominant cancer in young men, accounting for 60% of all malignant tumors in Caucasian men between the age of 25 and 45 years[27,28]. In general, these tumors have high cure rates. Nevertheless, treatment resistance is observed in at most 5% of patients[29,30]. The mechanisms of resistance are poorly understood, which is further elaborated in the following.

A brief introduction to TGCT

The formation of TGCT is initiated at early stages of embryonic development[28,29,31]. During development of the embryo, embryonic stem cells (ESC) are formed. A subset of ESC differentiates into organs, such as the skin and brain. Another, smaller subset of ESC is committed to the germ line[32]. These precursors of the germ cell lineage are referred to as primordial germ cells (PGC) and will further differentiate into spermatogonia in males[28,30]. This differentiation is an essential process, as a correct differentiation allows for the inheritance of genetic information in the next generation. However, in some cases, PCGs can be blocked in their differentiation, leading to the formation of TGCT[30,32]. The involvement of the germ line in heritability of genetic information is thought to contribute to the high cure rates observed for TGCT, but the exact mechanics are poorly understood[33].

In the first step of the development of TGCT, a precursor lesion (pre-GCNIS) is formed through polyploidization (Fig. 2B)[30]. Pre-GCNIS further accumulates copy number alterations and develops into GCNIS (also known as CIS). To be consistent with existing literature, we will use the abbreviation CIS in the rest of this text. CIS cells are initially dormant until puberty. It takes around 10-20 years after puberty for the CIS cells to further differentiate, accumulate somatic mutations and become invasive[30].

CIS cells can further develop into two subtypes of TGCT, seminoma and non-seminoma. Seminoma are homogeneous tumors and resemble PGC[27]. Non-seminoma are typically highly heterogeneous tumors and can differentiate into multiple types of tissue[27].

Non-seminomatous type 2 TGCT are heterogeneous tumors

Non-seminoma can consist of multiple differentiations (also called subtypes), including embryonal carcinoma (EC), teratoma (TE), yolk sac tumor (YST) and choriocarcinoma (CH)[27,28]. EC are pluripotent cells which are formed through an (epi)genetic reprogramming of CIS cells[30–32]. EC are able to further differentiate into TE, YST and CH[28,30–32]. In addition, embryoid bodies (EB) may form in the tumor, which resemble early embryonic structures[32]. Eventually, metastases can be formed. Throughout the development of the tumor, copy number alterations and somatic mutations accumulate. However, which mutations and alterations occur at which steps of the development of the tumor is poorly understood. It is currently not well defined which changes to the genomic constitution are required for tumor development, growth and the acquisition of resistance to (chemo)therapy. Using the method presented here, we aim to reveal a subclonal evolution tree for 2 cases of non-seminomatous TGCT with intrinsic resistance to treatment.

In the following, we describe of the model of TargetClone, the simulated data and TGCT datasets used to validate the method. Results on a simulated dataset and the TGCT data are shown in Section 3. Section 4 concludes the article with a discussion and an overview of future work.

2 Methods

In this section, we present the TargetClone method for inferring subclones and frequencies from targeted sequencing data. Before proposing the model, some basic definitions are introduced.

2.1 Basic definitions

First, we assume that tumors consist of multiple subclones with unique genotypes. We further assume that several regions of a tumor have been sampled, where each sample is a mixture of one tumor subclone and is potentially contaminated with normal cells. The frequency of the subclone and normal cells in the sample are denoted as $\vec{\mu}$ as defined in Eq. 1. As samples are assumed to at most consist of two components, m is fixed at 2. Thus, μ_1 indicates the ratio of normal cells in the same, and μ_2 indicates the ratio of tumor cells.

Second, we assume that the genomes of each of the samples have been segmented. A segment is defined as a chromosomal region across which the copy number and ratio of the parental alleles is the same, thus having the same lesser allele frequency at all measurement positions within the region (Fig. 4A). In Fig 4A, we show sample 1 and sample 4 from Fig. 3. However, now two additional chromosomal regions (the first two), or segments, are added to the cells. Furthermore, it is required that the segmentation is the same for all samples of the tumor. Thus, if the introduction of a new segment is only required in one sample, we will treat the region as two segments in other samples as well, even if this segmentation may not be necessary in the other samples. We further assume that a new segment always begins at the start of a chromosome. Every segment may contain p lesser allele frequency measurements.

We define the sample ploidy as $\vec{C}$ as in Eq. 2, where every chromosomal region i is a segment.

The measurements on segments can be handled in different ways

Each segment i may contain multiple lesser allele frequency measurements. As we assume that the segmentation has been performed such that every segment corresponds to one uniform copy number, it is possible to take the mean or median of all lesser allele frequency measurements on a segment to reduce computational time. Taking the mean or median of a segment can reduce the influence of artifacts introduced by the presence of noise in the measurements. However, taking the mean or median may also reduce resolution. Therefore, we also allow the method to find the best $\vec{C}$ and $\vec{\mu}$ based on all measurements on a segment. From our simulations (Section 2.3) we concluded that using the median results in the overall highest accuracy. For all results presented in the main paper, the median was used to obtain one measurement per segment. Details on the accuracies obtained for the other two metrics can be found in Section B.3 of the Supplementary Data.

2.2 TargetClone model

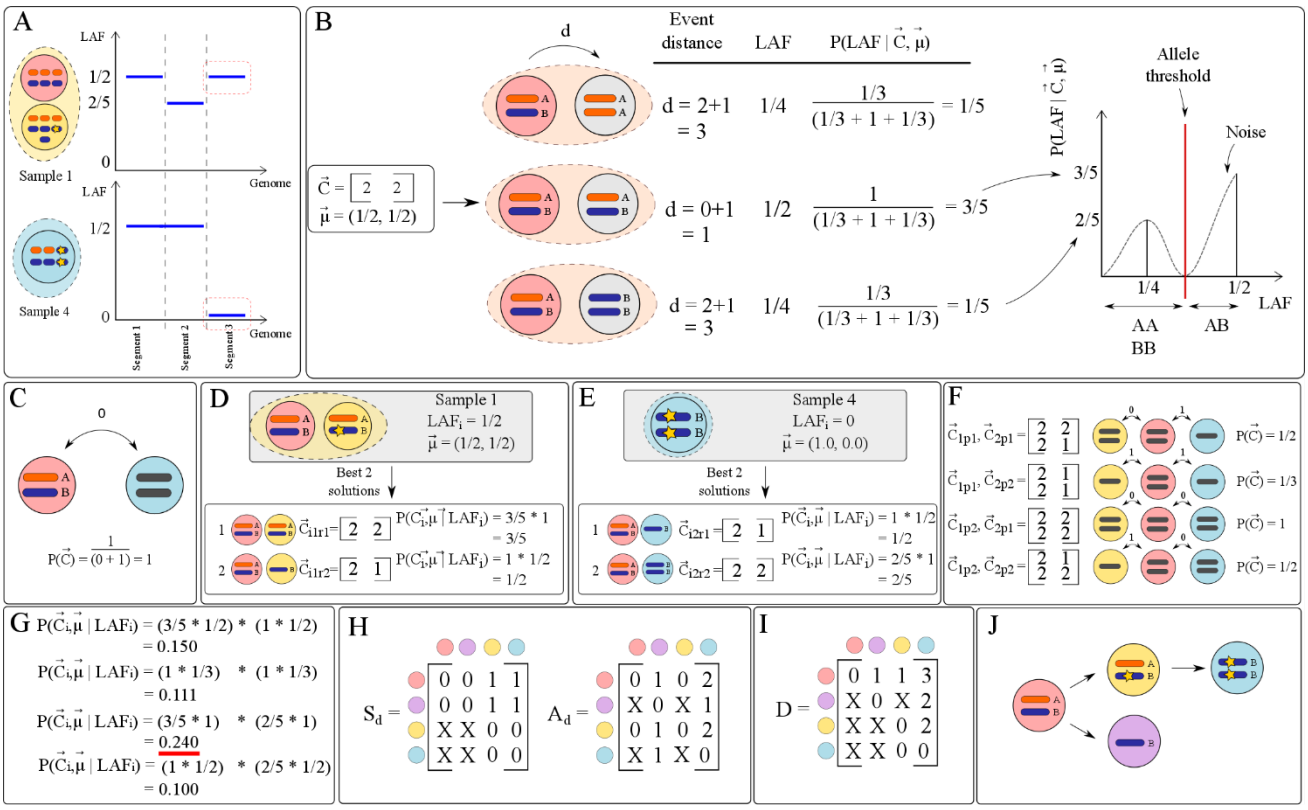

Fig. 4. Illustration of the methodology of TargetClone. All panels are discussed in the main text.

The aim of TargetClone is to find a combination of $\vec{C}$ and $\vec{\mu}$ per tumor sample that maximizes the probability of observing the lesser allele frequency measurements of that sample (which is defined as $\overrightarrow{LAF}$):

$$\underset{\vec{C},\vec{\mu}}{arg\,max}\ \mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{LAF}\right) \tag{4}$$

We assume that $\mathbb{P}(\overrightarrow{LAF})$ is only affected by sequencing noise and is therefore fixed. Using Bayes rule, we can rewrite Eq. 4 as (for the full derivation see Section A of the Supplementary data)

$$\underset{\vec{C},\vec{\mu}}{arg\,max}\ \mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{LAF}\right) = \mathbb{P}\left(\overrightarrow{LAF}\middle|\vec{C},\vec{\mu}\right)\mathbb{P}\left(\vec{C}\right) \tag{5}$$

First, we will describe how $\mathbb{P}\left(\overrightarrow{LAF}\middle|\vec{C},\vec{\mu}\right)$ is computed. Then, we explain the computation of $\mathbb{P}\left(\vec{C}\right)$. Third, we give a description of how Eq. 5 is maximized. Fourth, as an additional step, we explain how we infer the most likely alleles of the tumor subclone given $\vec{C}$ and $\vec{\mu}$. Finally, we end with how the subclonal evolution trees are reconstructed for the tumor samples.

2.2.1 Computing $\mathbb{P}\left(\overrightarrow{\mathbf{LAF}}\middle|\vec{\mathbf{C}},\vec{\mu}\right)$

To compute the probability of observing lesser allele frequency measurements in a sample given any combination of $\vec{C}$ and $\vec{\mu}$, we first assume that all segments are independent. This assumption is reasonable as we have defined a segment as a chromosomal region with the same copy number and ratio of parental alleles. Therefore, the copy number of one segment does not depend on the copy number of another segment.

Per segment i, we can thus aim to infer a copy number for the tumor subclone at $\vec{C}_{i,2}$ that together with $\vec{\mu}$ maximizes $\mathbb{P}\left(LAF_i|\vec{C}_i,\vec{\mu}\right)$. As discussed previously in the Introduction, inferring the most likely copy number for the tumor subclone is not easy as some lesser allele frequencies can be explained by multiple copy numbers. We there made the assumption that such copy number ambiguities can potentially be resolved by incorporating information about the possible copy numbers of other samples. As tumors develop as part of an evolutionary process, the number of genomic changes that occurred between all subclones in the tumor is assumed to be minimal. Thus, if a set of lesser allele frequency measurements on a segment can be explained by multiple $\vec{C}_i$, the likelihood of observing that set of lesser allele frequency measurements is expected to be higher for a $\vec{C}_i$ for which the number of changes that need to be made to obtain the copy numbers at the same segment in the other tumor subclones is small than for a $\vec{C}_i$ for which this number of changes is large.

Furthermore, the assumption of the existence of an evolutionary process implies some biological restrictions to the possible copy numbers that a tumor subclone can have at a given segment. For example, if the parent of a tumor subclone in a sample has a copy number of 0, then the tumor subclone itself can never have any other copy number than 0. We can make this idea more specific by reasoning on the level of alleles, which is discussed in the next section.

Copy number ambiguities may be resolved by using alleles

Every copy number at a segment in the tumor subclone of the sample can potentially have a different combination of alleles. In the left part of Fig. 4B, we show dotted circles that indicate a sample with one normal cell and one tumor subclone where $\vec{\mu}$ = (1/2, 1/2). Every sample shows the different possible combination of alleles that a tumor subclone can have at a single

segment, in this particular example for a copy number of 2. This copy number can be obtained with the allele combinations AA, BB and AB.

We define an event as the gain or loss of an allele. Let the event distance be a metric that indicates how many alleles have been lost or gained between two cells. For example, the event distance at a chromosomal region from the normal cell with alleles AB to a tumor subclone with alleles BB is 2 (sample 3 in Fig. 4B), corresponding to the loss of allele A and the gain of allele B. Note that this event distance is asymmetrical. For example, there exists no valid event distance from the normal cell back to the tumor subclone. The tumor subclone has lost allele A, which can never be regained. Thus, if we knew which subclone is the precursor of the tumor subclone in each sample, and if we knew the exact combination of alleles at every segment in these other subclones, it would be possible to resolve copy number ambiguities by selecting the copy numbers for corresponding alleles that minimize the event distance between all subclones. However, at this stage of the method we lack this information as we do not have copy number information. The only alleles that are known are the alleles of normal cells, which are the alleles AB.

Therefore, we initially aim to resolve copy number ambiguities by assigning a higher probability to lesser allele frequency measurements when the corresponding alleles for the given copy number have a smaller event distance to AB.

The steps required to compute the probability of observing lesser allele frequency measurements in a sample given a $\vec{C}$ and $\vec{\mu}$ are discussed below.

Step 1: computing the possible lesser allele frequencies given $\vec{C}$ and $\vec{\mu}$

For a given $\vec{C}$ and $\vec{\mu}$, we first determine the possible alleles that can be present in the tumor subclone given the copy numbers defined in $\vec{C}$ per segment (Fig 4B, samples). For example, if $\vec{C}_i$ is [2 2], the possible alleles in the tumor subclone can be AA, AB or BB. For each possible set of alleles, we first compute the lesser allele frequency that the scenario will generate using Eq. 3. This computation is repeated for every segment independently.

Step 2: similarity between a tumor subclone and the normal component

Following step 1, we compute the similarity between the alleles of a normal cell and any possible combination of alleles for the tumor subclone. First, the event distance is computed from the normal cell to the tumor subclone per allele combination (Fig. 4B). This event distance is increased by one to prevent divisions by zero in later computation. Following the computation of the event distance, a similarity score is computed as the normalized event distance given the event distances of all other possible combinations of alleles q that can be made with this $\vec{C}_i$ and $\vec{\mu}$. The similarity score s between the normal cell a and the tumor subclone with a specific combination of alleles b is thus computed as

$$s = \frac{\frac{1}{event\ distance\ (a,b)+1}}{\sum_{l=1}^{q}\frac{1}{event\ distance\ (a,l)+1}} \quad (6)$$

Formally, similarity scores do not equal probabilities. However, we reason that the similarity score can be used as a probability as the score is proportional to actual probabilities.

Step 3: computing the probability of measuring lesser allele frequencies given $\vec{C}$ and $\vec{\mu}$

In the final step, we assign a probability to the measured lesser allele frequencies in a sample given $\vec{C}$ and $\vec{\mu}$ by per segment mapping the lesser allele frequencies that each possible combination of alleles in the tumor subclone can generate to the corresponding similarity scores. For any lesser allele frequency that can be measured with multiple combinations, we sum the similarity scores of the combinations. The result of this final step is a discrete probability distribution of lesser allele frequencies measurements that can be obtained with the given $\vec{C}$ and $\vec{\mu}$. We can read the probability of any lesser allele frequency measurement from this distribution. In practice, measured lesser allele frequencies will never equal the true values as a result of sequencing noise. Therefore, we also add noise to the probability distributions.

Adding sequencing noise to the model

We assume that sequencing noise follows a normal distribution with standard deviation σ. We further assume that the noise is the same for every lesser allele frequency measurement. The probability density function is modeled as a mixture of Gaussians where the number of components equals the number of unique lesser allele frequencies that can be measured with a $\vec{C}_i$ and $\vec{\mu}$ (Fig. 4B, distribution). For segment i, this distribution is $\mathbb{P}\left(\overrightarrow{LAF_i}\middle|\vec{C}_i,\vec{\mu}\right)$. The means of the components equal the corresponding lesser allele frequency measurements. σ is measured as the standard deviation in the reference sample, which is approximated to be 0.02. Finally, under the assumption that all segments are independent, the $\mathbb{P}\left(\overrightarrow{LAF}\middle|\vec{C},\vec{\mu}\right)$ for n segments is computed as

$$\mathbb{P}\left(\overrightarrow{LAF}\middle|\vec{C},\vec{\mu}\right) = \prod_{i=1}^{n}\mathbb{P}\left(\overrightarrow{LAF_i}\middle|\vec{C}_i,\vec{\mu}\right) \quad (7)$$

2.2.2 Computing $\mathbb{P}(\vec{C})$

As stated previously when introducing the event distance, we assume that there exist evolutionary relationships between the samples and that the number of events between subclones is minimal. Therefore, the copy numbers of segments are also expected to be similar across subclones. The likelihood of a $\vec{C}$ is higher when the total distance between the copy numbers of all subclones is minimal. The distance d_s between subclones is defined as

$$d_s = \left(\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left|\vec{C}_i - \vec{C}_j\right|\right) + 1 \quad (8)$$

As explained before, the distance between subclones is not immediately a probability, but is proportional to the probability. Like before, we take the reciprocal of d_s to convert the distance to a similarity score. This score is $\mathbb{P}(\vec{C})$ (Fig. 4C). In the example of Fig. 4C, the blue cell indicates a tumor subclone. As the alleles of the tumor subclone are not known at this point, the alleles are omitted from the figure.

2.2.3 Maximizing $\mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{LAF}\right)$

We aim to find the combination of $\vec{C}$ and $\vec{\mu}$ per sample that maximizes the likelihood of our initial lesser allele frequency measurements in each sample. The assumption is made that all segments are independent. Therefore, for any $\vec{\mu}$, we can find the most likely $\vec{C}$ per segment. Segments typically consist of multiple lesser allele frequency measurements. We first give a description of how the method can handle multiple measurements per segment. Then, we continue with describing how we maximize $\mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{LAF}\right)$.

$\mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{LAF}\right)$ is maximized with an exhaustive search

As the most likely $\vec{C}$ is computationally easy to obtain given an initial $\vec{\mu}$, we first perform a greedy exhaustive search where we iteratively search through all possible $\vec{\mu}$, where the frequencies are increased with a step size of 0.01 (details on the choice of step size are discussed in Section B.2 of the Supplementary data).

Given any $\vec{\mu}$, we compute $\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right)$ for any $\vec{C}$ where the copy number of the tumor component lies between *kmin* and *kmax*. By default, *kmin* is set to 0 and *kmax* is set to 6. This computation is performed for each segment independently as

$$\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right) = \prod_i^n \mathbb{P}\left(\overrightarrow{LAF_i} \middle| \vec{C}, \vec{\mu}\right) \mathbb{P}\left(\vec{C_i}\right) \quad (9)$$

A semi-greedy approach is used to include information across samples
For any $\vec{\mu}$, we obtain the l $\vec{C}$ with the highest $\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right)$. Here, we fix l at 2. This is shown in Fig. 4D and Fig. 4E. In Fig. 4D, we show the two best $\vec{C}_i$ for sample one. The best solution has a copy number of two in the tumor subclone, whereas the second best solution has a copy number of one in the tumor subclone. $\vec{C}_{i1r1}$ indicates that this $\vec{C}_i$ is for sample one and has rank one. $\mathbb{P}\left(\overrightarrow{LAF_i} \middle| \vec{C_i}, \vec{\mu}\right) = 3/5$ is obtained for a copy number of 2 with a lesser allele frequency measurement of 0.5 based on the probability distribution shown in Fig. 4B. $\mathbb{P}\left(\vec{C_i}\right) = 1$ is computed as shown in Fig. 4C. For sample 4, the best 2 $\vec{C}_i$ are similarly computed. As we see, the best $\vec{C}_i$ is no the correct solution, but the second best $\vec{C}_i$ is. Thus, we here show that it is useful to include the second best $\vec{C}_i$ in our computations compared to the greedy approach.

For each segment in every sample, we make all possible combinations with the copy numbers of the tumor components across all samples. An example of this step for the samples of Fig. 4D and Fig. 4E is shown in Fig. 4F (left). A total of 4 combinations can be made for these samples. For every possible combination, we re-compute $\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right)$ per sample as defined in equation 3 (Fig. 4G), but $\mathbb{P}\left(\vec{C_i}\right)$ is now computed using the copy numbers of all samples (Fig. 4F, right). The $\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right)$ of the two samples are multiplied. Thus, a higher probability is assigned to combinations where the copy numbers do not deviate much between the samples at a specific segment. From Fig. 4G we observe that the third combination has the highest likelihood given our lesser allele frequency measurements in the two samples. This combination is indeed the desired solution. In Section 3.2 of the Results we demonstrate the benefits of including information across samples (semi-greedy) to compute $\mathbb{P}\left(\vec{C}, \vec{\mu} \middle| \overrightarrow{LAF}\right)$ as compared to when $\vec{C}$ is selected based on only the information within each sample individually (greedy) for our simulation data.

2.2.4 Inferring the most likely alleles of the tumor subclones

As described in Section 2.2.1, $\vec{C_i}$ can correspond to multiple possible combinations of alleles (Fig. 4B). The distribution $\mathbb{P}\left(\overrightarrow{LAF_i} \middle| \vec{C_i}, \vec{\mu}\right)$ explains how likely it is to measure the lesser allele frequencies that correspond to the combinations of alleles. Thus, in this distribution, each region between one valley to the next represents the probability of measuring one lesser allele frequency given $\vec{C_i}$ and $\vec{\mu}$. The regions are shown in the distribution of Fig. 4B, where the allele threshold indicates a switch from one valley to the next. For example, if we consider the distribution in Fig. 4B, the first valley until the second corresponds to a lesser allele frequency of 1/4, and the alleles can be AA or BB. Thus, if our actual lesser allele frequency measurement in the sample at this segment is 1/4, then the alleles will be either AA or BB given this $\vec{C_i}$ and $\vec{\mu}$. As it is not possible for us to distinguish between AA or BB based on the lesser allele frequency measurements alone, we always arbitrarily select the solution with the most frequent number of B alleles by default. In our example, we would thus select the alleles BB for the tumor subclone at this specific segment. The alleles of the tumor subclone at all segments are stored in $\vec{A}$.

2.2.5 Reconstructing the subclonal evolution tree

To reconstruct subclonal evolution trees for tumor subclones, a distance matrix is defined which contains the distances between the subclones. This distance matrix can be computed in various ways. We can for example compute the distances between subclones based on only $\vec{C}$, or only $\vec{A}$. We expect that making combinations with different types of information can improve the reconstruction of subclonal evolution trees. For example, alleles are good indicators of which relationships may be possible between subclones. If allele A or B has been lost in one subclone, then it can never be regained. For copy numbers, the same idea holds for having zero copies, where the copy number can never again increase. Somatic variants in a tumor subclone are also good indicators of relationships between subclones. By definition of the infinite sites assumption, a somatic variant is not lost once gained. Therefore, a subclone without a somatic variant at a specific location can never be the child of a subclone that does contain a somatic variant at that location. These restrictions are useful in reducing the possible number of subclonal evolution trees that can be made for a set of tumor subclones. However, especially when a limited number of somatic variants have been measured, the somatic variants alone may not always provide enough information to accurately reconstruct subclonal evolution trees. However, this information may be present in for example the alleles or copy numbers. Thus, we expect to observe that making combinations between $\vec{C}$, $\vec{A}$ and somatic variants to compute distance matrices can increase reconstruction accuracy of subclonal evolution trees.

The distance matrix based on only $\vec{C}$ is determined by computing d_s (Section 2.2.2) between all subclones. The distance matrix based on only $\vec{A}$ is computed similarly, but is instead based on the sum of the event distances for all segments between subclones. An example of a distance matrix based on $\vec{A}$ is shown in Fig. 4H (the matrix is called A_d). The distance matrices in this figure are computed for the subclones shown in Fig. 3. An 'X' in the matrix in the figure indicates that the relation between subclones is not possible. In the actual distance matrix, an 'X' is encoded as an infinite distance. The distance matrix based on only somatic variants is computed slightly differently. For all somatic variants that have been measured in all samples, we use an indicator value of 1 to indicate that the somatic variant is present in the subclone (variant allele frequency > 0), otherwise the indicator value is 0. Using these binary indicator values, we define the distance matrix between subclones based on somatic variants as follows. Between any subclone, if no somatic variant is lost, the distance is 1. Otherwise, the distance is infinite. An example of such a distance matrix for the somatic variants is shown in Fig. 4H (called S_d).

The distance matrix of the somatic variants can be combined with the distance matrices of $\vec{C}$ and $\vec{A}$. This combination is done through weighing the $\vec{C}$ and/or $\vec{A}$ matrices with the distance matrix of the somatic variants by summing the matrices (Fig. 4I shows a distance matrix when the distance matrix for $\vec{A}$ is summed with the distance matrix of the somatic variants).

The (weighted) distance matrix is used as input to Edmond's algorithm to reconstruct a subclonal evolution tree[34]. Edmond's algorithm infers a spanning arborescence, which is an acyclic directed graph, from a given distance matrix such that the distances on all branches in the tree sum up to a minimum value.

2.3 Simulation data

To generate our simulated dataset, we first generated 100 artificial tumors. Then, we took samples of the artificial tumors with which we validated the

method. Both steps are elaborated on in their respective sections below. Additional details can be found in section B.1 of the Supplementary Data.

Step 1: artificial tumor generation

An ensemble of 100 datasets was created, each resembling an artificial tumor bulk containing 5 subclones each. For every dataset individually, we started by randomly introducing 500 heterozygous SNP positions into a normal, diploid genome at which the lesser allele frequency will be measured. In addition, 10 positions were defined at which somatic variants will be measured. These SNP positions were non-uniformly assigned to a total of 35 segments. Starting from the normal genome, we generated a new subclone by probabilistically introducing 5 rounds of allelic copy number variations into the normal genome, where each genomic event affects an entire segment. Additionally, a random number (between 0 and 10) of somatic variants were added to the subclone. We define the presence and absence of a somatic variant in a subclone with a 1 or 0, respectively. Somatic variants are allowed to be gained, but never lost. Furthermore, somatic variants are not allowed to be introduced at the same position.

As allelic copy number variations are theoretically reversible, the copy number events are allowed to occur with replacement. Both the normal component and the new subclone were added to the artificial tumor, randomly selecting their frequencies such that the total frequency inside the artificial tumor sums to 1. In the next step, more subclones were generated by probabilistically determining the parent based on the frequencies of the existing subclones. Again, each subclone was subjected to 5 rounds of introducing genomic events and adding somatic variants. This iterative process was repeated until the desired number of 5 subclones was reached.

Step 2: sampling from the artificial tumors

We define a sample as a mixture of one subclone and normal cells. Each subclone was assigned to a sample exactly once without replacement, generating a total of 5 samples. We added a random percentage of normal contamination to each sample, after which the lesser allele frequency measurement was computed at every segment. The lesser allele frequency measurement values at the additional SNP positions at each segment were determined by sampling from a normal distribution truncated at a lesser allele frequency measurement of 0.5 where the mean equals the lesser allele frequency measurement of the segment and the $\sigma = \{0, 0.02, 0.04, 0.06, 0.08, 1\}$. The final lesser allele frequency measurement assigned to each segment is the median of the lesser allele frequency measurements at all SNP positions located on that segment.

2.4 Real data

The method was applied to 2 cases of chemotherapy-resistant testicular germ cell tumors. Both cases were subjected to IonTorrent sequencing. The first case (T6107) consists of 15 samples, in which the lesser allele frequency was measured at 427 heterozygous SNP positions. The variant allele frequency was measured for 14 somatic variants. The genomes were manually segmented into 35 segments based on visual inspection by an expert. The second case (T3209) consists of 25 samples, in which the lesser allele frequency was measured at 431 heterozygous SNP positions. The variant allele frequency was measured for 31 somatic variants. The genomes were again manually segmented into 31 segments based on visual inspection by an expert. The measurements in the reference samples have a standard deviation of approximately 0.02, which we use as an estimate of the sequencing noise in both cases.

More details on these data are described in Section C of the Supplementary Data.

Measuring the accuracy

The accuracy of the inferred $\hat{\vec{C}}$ across n segments compared to the true $\vec{C}$ is computed as

$$1 - \frac{1}{n}\sum_{i=1}^{n} |\vec{C}_i - \hat{\vec{C}}_i| \quad (10)$$

The accuracy of the inferred $\hat{\vec{\mu}}$ compared to the true $\vec{\mu}$ is similarly computed as:

$$1 - \frac{1}{m}\sum_{i=1}^{m} |\vec{\mu}_i - \hat{\vec{\mu}}_i| \quad (11)$$

The accuracy of the inferred $\hat{\vec{A}}$ compared to the true $\vec{A}$ is computed as:

$$1 - \frac{1}{n}\sum_{i=1}^{n} |\vec{A}_i - \hat{\vec{A}}_i| \quad (12)$$

Additionally, it is interesting to observe how often copy number ambiguities are present in our simulation set. To see if our method can resolve copy number ambiguities, we measured an *ambiguity score*. This score measures how often we in a sample infer an incorrect copy number at a segment, but the lesser allele frequency at that segment remains the same compared to the true subclone.

Accuracy of reconstructing the subclonal evolution trees

To measure how well the reconstructed subclonal evolution trees match the true trees, we define an accuracy score as follows. For every subclone, we compare if the correct parent has been inferred for this subclone. Let I_p be an indicator variable that is 1 when the correct parent has been inferred. The accuracy is then computed as

$$1 - \frac{1}{m}\sum_{i=1}^{m} I_p \quad (13)$$

3 Results

Here, we present the results of TargetClone on simulation data and a case of 2 testicular germ cell tumors with intrinsic resistance to chemotherapy. First, we will discuss how well the method can infer $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ and subclonal evolution trees for our simulation data in Section 3.1. Then, we show the benefits of determining the most likely $\vec{C}$ by utilizing the possible copy number solutions across samples in Section 3.2. Finally, we discuss the reconstructed subclonal evolution trees for the real data in Section 3.3.

3.1 Simulation data

We started with inferring $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ for the 100 datasets in our simulation dataset, while increasing the noise levels from 0 to 0.1 standard deviations as defined in Section 2.3. The results shown here correspond to the specific scenario where we infer the most likely $\vec{C}$ and $\vec{A}$ from the median lesser allele frequency measurement of each segment. The accuracy of how well $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ are reconstructed at each noise level is shown in Fig. 5. Fig. 5A focuses on how well $\vec{C}$ is reconstructed in our simulation data. We notice a couple of things from this result. First of all, the reconstruction accuracy of $\vec{C}$ drops as the noise level increases. There is little difference in reconstruction accuracy between noise standard deviations of 0 and 0.02, with a low spread in the accuracy values. We observe that the median reconstruction accuracy decreases to approximately 0.2 at a noise level of 0.1 standard deviations.

Second, interestingly, the ambiguity score (blue boxplots) remains at a median accuracy of approximately 0.95 as the noise increases.

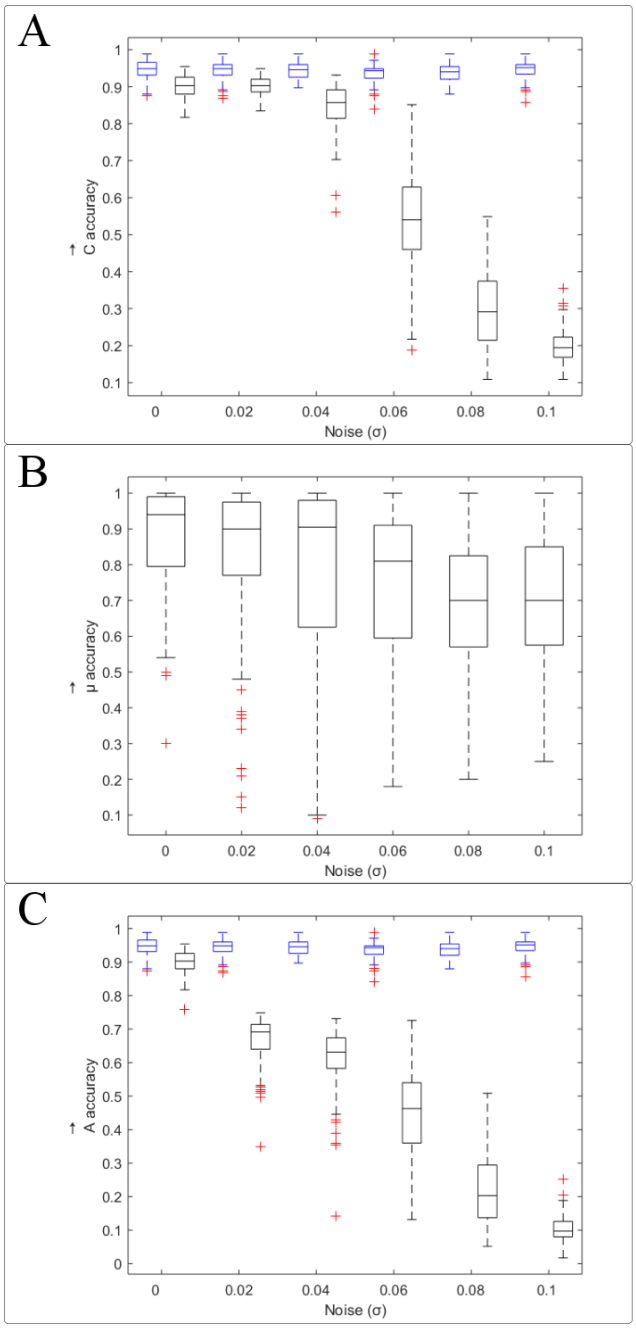

Fig. 5. Accuracy of inferring (A) $\vec{C}$, (B) $\vec{\mu}$ and (C) $\vec{A}$. The results are described in the text.

As discussed previously, the ambiguity score is a measure of on how many segments $\vec{C}$ is incorrect in the inferred tumor subclone, but the lesser allele frequency measurement that is obtained for the incorrectly inferred copy numbers at those segments is the same as in the simulated tumor subclone. What the ambiguity score reveals in these results is that despite that the inferred $\vec{C}$ deviates from the true $\vec{C}$ in on average 10% of all segments in the samples per simulated dataset, the lesser allele frequency measurements at those segments are the same as in the simulated samples, even as the noise increases. Thus, the inferred $\vec{C}$ may deviate from the ground truth, but the inferred copy numbers are ambiguous copy numbers. As a reminder, these ambiguous copy numbers can for example be 0, 2 or 4, which each generate a lesser allele frequency of 1/2 when $\vec{\mu}$ = (1/2, 1/2) in the simulated sample. From this observation we learn that our method has more difficulties with resolving copy number ambiguities when the noise levels increase.

Third, we observe that even in the scenario where no noise has been added to the lesser allele frequency measurements, the median reconstruction accuracy is 0.9. This value indicates that in all samples, approximately 10% of all inferred copy numbers in $\vec{C}$ is different from the ground truth. 5% of the inferred copy numbers that are different from the ground truth are ambiguous copy numbers that have the same lesser allele frequency as the simulated samples. The other 5% is explained by segments in the simulated samples at which the method infers copy numbers different from the ground truth that are also not ambiguous for the input lesser allele frequency measurements, and can be viewed as a true error.

Fig. 5B shows the accuracy with which $\vec{\mu}$ is inferred. It can be observed from this figure that the accuracy decreases slower than for the inference of $\vec{C}$ as the noise in the lesser allele frequency measurements increases. At the highest noise level, the median reconstruction accuracy does not decrease below 0.7. The median reconstruction error rate at a noise level of 0 is approximately 5%. As discussed above, the reconstruction error of $\vec{C}$ is approximately 10% in the scenario without noise. From these error rates alone, it is not possible to tell if for 5% of the samples for which the inferred $\vec{C}$ deviates from the true $\vec{C}$, the inferred $\vec{\mu}$ of that same sample is also different from the true $\vec{\mu}$.

Overall, we observe that the spread in the accuracy of inferring $\vec{\mu}$ is relatively large at all noise levels compared to the accuracy of inferring $\vec{C}$ and thus varies largely between samples.

Fig 5C shows the reconstruction accuracy of $\vec{A}$. Similar to the accuracy of inferring $\vec{C}$, the accuracy with which $\vec{A}$ is reconstructed decreases as the noise level increases. However, the reconstruction accuracy of $\vec{A}$ decreases faster starting from a noise level of 0.02. The explanation for this behavior lies with the $\mathbb{P}\left(\overrightarrow{LAF_i}\middle|\vec{C_i},\vec{\mu}\right)$ distribution that we use to determine the most likely alleles given a lesser allele frequency measurement. For any given $\vec{C}$ and $\vec{\mu}$, we determine the alleles by determining which noise-less lesser allele frequency is most likely associated with our measurement that is affected by noise by finding the nearest peak in the distribution. Thus, even if the inferred $\vec{C}$ may be correct, due to the presence of noise in the lesser allele frequency measurements the nearest peak could be a different one from expected, which results in the inference of the incorrect alleles.

Conclusions

As discussed above, we observed from our results on the simulation data that $\vec{\mu}$ is in general more difficult to estimate than $\vec{C}$ and $\vec{A}$, which can be observed from the larger spread in the accuracies of inferring $\vec{\mu}$ compared to $\vec{C}$ and $\vec{A}$ (Fig. 5B). One possible explanation for this behavior is that inferring $\vec{\mu}$ relies on the lesser allele frequency measurements at all segments in a sample. Therefore, if the inferred copy number is incorrect at one segment of the sample, it may influence the most likely $\vec{\mu}$ which in combination with the full $\vec{C}$ results in the highest $\mathbb{P}\left(\vec{C},\vec{\mu}\middle|\overrightarrow{\mathrm{LAF}}\right)$.

Furthermore, we observed in Fig. 5A that a decrease in the reconstruction accuracy of $\vec{C}$ is affected by ambiguous copy numbers that result in the same lesser allele frequency. As our method on average never scores accuracies of inferring $\vec{C}$ above the ambiguity score, it is thus not possible to resolve all ambiguities even when information is included across samples.

Finally, it is more difficult to accurately infer $\vec{A}$ than $\vec{C}$. The inference of $\vec{A}$ is especially influenced by noise and already results in a median reconstruction error of approximately 31% at a noise level of 0.02 standard

deviations. Therefore, the inferred alleles ought to be interpreted with care in real data where sequencing noise in the lesser allele frequency measurements is not absent.

Reconstructing subclonal evolution trees

For each of our simulation datasets, we reconstructed subclonal evolution trees for the 5 subclones using only $\vec{C}$, only $\vec{A}$, only somatic variants, combining $\vec{C}$ with somatic variants, and combining $\vec{A}$ with somatic variants. As described previously, we expect that combining information can help in improving the accuracy with which subclonal evolution trees are reconstructed. In Fig. 6, we show an example of the true subclonal evolution tree for an arbitrary simulation dataset (Fig. 6A) compared to the inferred tree by the model for that particular dataset (Fig. 6B). The tree was generated using a distance matrix based on the combination of $\vec{C}$ and somatic variants. In this figure, we see that the inferred subclonal evolution tree is different from the inferred subclonal evolution tree. The reason that the inferred tree is different lies with the underlying simulated measurements. Both subclone 3 and 4 have the same somatic variants. The true copy numbers in subclone 3 and subclone 4 only differ on one segment, whereas both subclones are highly similar to a normal cell (subclone 1). The distance from subclone 3 to subclone 1 is smaller than the distance from subclone 3 to subclone 4. Thus, as neither the copy numbers or somatic variants are enough to fully infer the parents of all subclones, the smaller distance from subclone 4 to subclone 1 rather than from subclone 4 to subclone 3 results in an incorrectly inferred tree. Based on the previously described accuracy measurements for trees, the inferred tree would receive an accuracy value of 0.8 as the parents of 4 out of 5 subclones are correctly inferred.

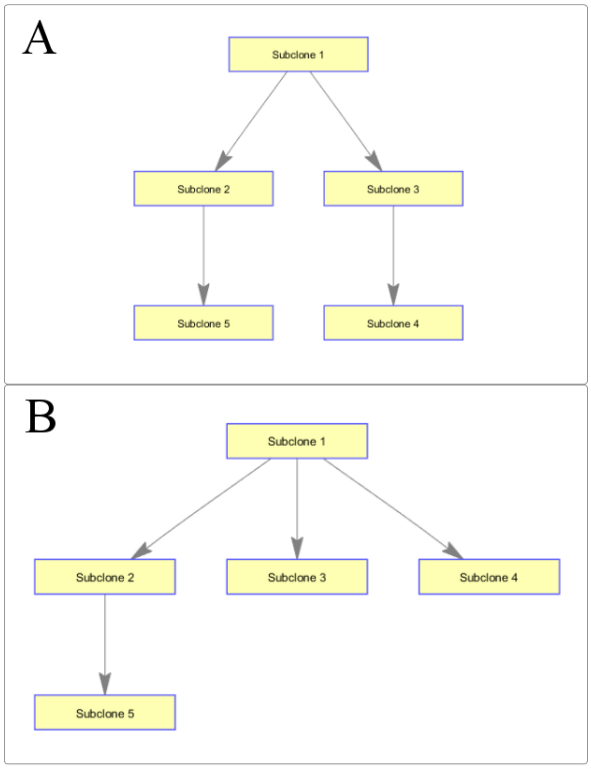

Fig. 6. (A) Artificial subclonal evolution tree and (B) an incorrectly inferred subclonal evolution tree for the same subclones.

In Fig. 7 the accuracy is shown with which we reconstruct the subclonal evolution trees in our simulation data. The results for the tree reconstructing using combinations of $\vec{C}$, $\vec{A}$ and somatic variants will be discussed in their respective sections below.

Using only $\vec{C}$

In Fig. 7A, we see that the median reconstruction accuracy when using only the ploidy is 0.4. Overall, an accuracy of 0.4 indicates that $\vec{C}$ alone is not enough to pinpoint relationships between samples. One important reason for this is that copy numbers alone do not always provide sufficient information on which relationships are (not) possible. For example, we know that when all copies are lost at a segment, the copy number can no longer increase. This allows us ensure that a sample with a homozygous deletion can never be the parent of a sample that lacks this homozygous deletion. However, losing all copies of a chromosome has an event distance of 2 if the parent is a normal cell. Therefore, based on the probabilistic introduction of allelic copy number changes into the simulated samples as described in Section 2.3, homozygous deletions are not common in our simulation data. Thus, relationships between samples are difficult to infer based on $\vec{C}$ alone in our simulation data.

Using only $\vec{A}$

From Fig. 7B we observe that the average clonal tree reconstruction accuracy based on $\vec{A}$ is approximately 0.35, which is very low. In Section 2.2.5, we described that we expected that using alleles can restrict possible relationships between samples and may improve reconstruction accuracy. From the results shown here, we see that this is not necessarily true. We observed a reconstruction accuracy of 0.9 for $\vec{A}$ in the scenario where no noise is added to the measurements. However, the median reconstruction accuracy of the clonal evolution tree never exceeds 0.8 in the same scenario. The low accuracy can be explained by the following.

The current method used to reconstruct the clonal evolution trees is very sensitive to mistakes. The distance computed is based on the sum of the distances between all segments across all samples. Due to the use of this distance matrix in reconstructing the clonal evolution trees, we are prone to making mistakes. For example, consider the scenario where a segment in $\vec{A}$ is inferred to have the alleles AA, while the simulated sample actually has alleles AB at that segment. If all the other samples at the same segment are inferred to have alleles AB, then the distance computed from the sample with AA to the samples with AB will be infinite and this relationship is restricted in the final reconstructed tree. However, the inference of AA instead of AB may be possible due to noise in the measurements and is erroneous (Fig. 4B, distribution). Therefore, in actuality, the relationship between the sample should have never been restricted. Thus, erroneously inferred alleles at only one segment can already cause the introduction of big errors in the final clonal evolution tree when using distance matrices containing distances between samples.

Using only somatic variants

Fig. 7C reveals that the median reconstruction accuracy for the clonal evolution trees is 0.6 when using only somatic variants, which is higher than the accuracies measured when using only $\vec{C}$ or $\vec{A}$. As the somatic variant information is not inferred by the model, but actually measured, it

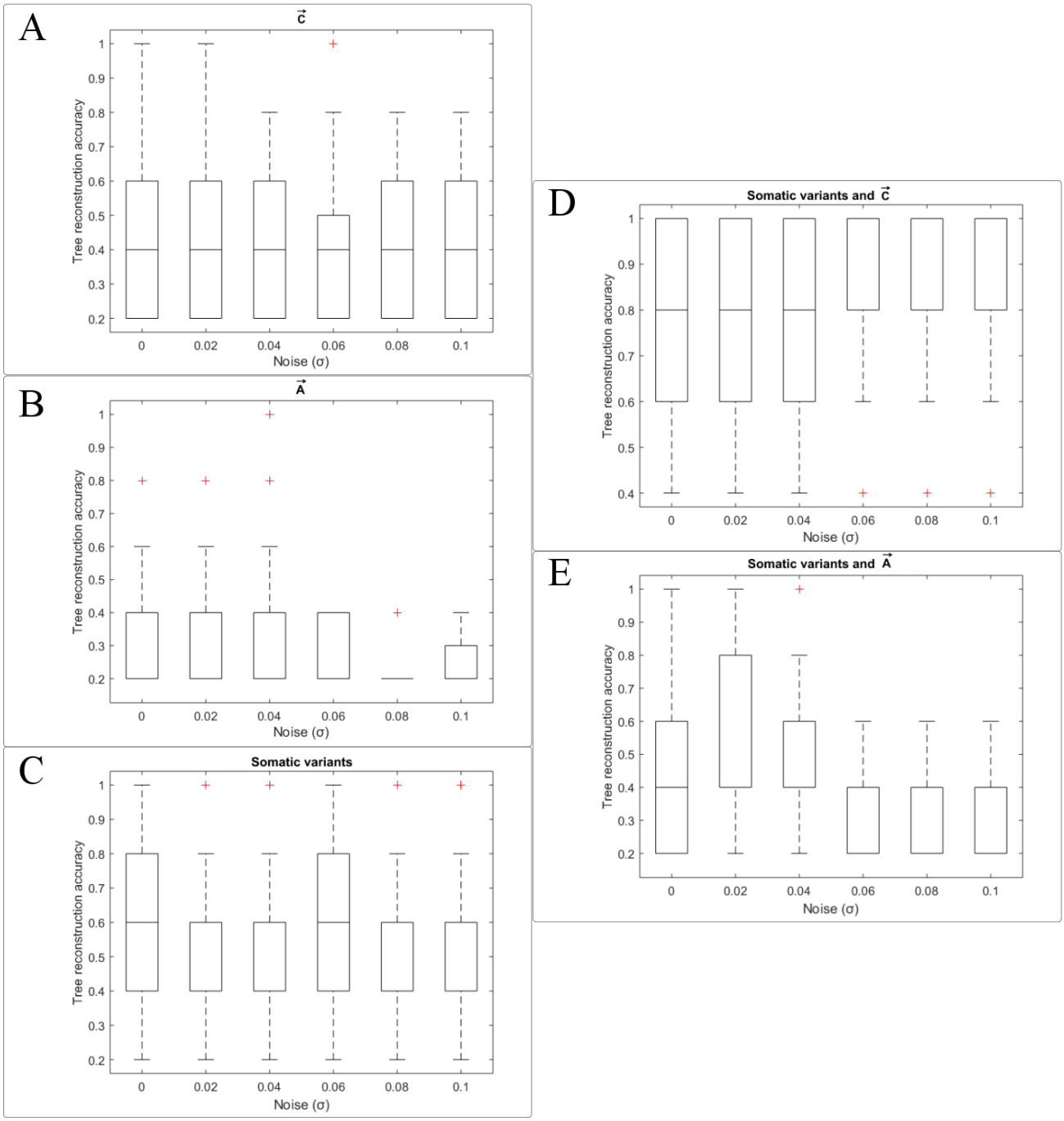

Fig. 7. Reconstruction accuracy of different combinations of $\vec{C}$, $\vec{A}$ and somatic variants. The panels are described in the text.

is not unexpected that this information results in higher accuracies. The somatic variants are introduced into the simulation dataset as binary values (present or absent). Therefore, we expect that there should not be large differences between the reconstruction accuracies across the different noise levels. The variation in accuracy that is observed is caused by the differences in the actual simulated data, as we generate 100 new datasets for each noise level.

Combining $\vec{C}$ with somatic variants

As somatic variants alone already have a median reconstruction accuracy of 0.6, it is not unexpected that the addition of $\vec{C}$ increases the average reconstruction accuracy to 0.8 (Fig. 7D).

Combining $\vec{A}$ with somatic variants

From Fig. 7E, we can observe that the average reconstruction accuracy is approximately 0.4. This is a decrease from the reconstruction accuracy of 0.6, which was obtained using somatic variants only. One would expect that if new information is added, the reconstruction accuracy will increase. However, due to the previously described issue introduced by erroneously

inferred alleles, it becomes a lot more difficult to accurately infer the clonal evolution tree. Even though the relations between samples may have been inferred based on the somatic variants with a median accuracy of 0.6, the relationships are affected by the distances computed based on the alleles when weighing the matrices. If there exist incorrect distances in the distance matrix computed based on $\vec{A}$ due to errors in this matrix, the weighted distance matrix will also be incorrect and lead to the reconstruction of incorrect clonal evolution trees.

Conclusions

If we compare the subclonal tree reconstruction accuracies for all combinations of $\vec{C}$, $\vec{A}$ and somatic variants, we observe that the highest median reconstruction accuracy is obtained when combining $\vec{C}$ with somatic variants, followed by using somatic variants only. Furthermore, we observe that reconstructing the clonal evolution trees is robust to noise. Thus, we are able to correctly infer the parent of a subclone with an accuracy between 0.6 and 1 at a noise level of 0.1, where the median reconstruction accuracy of $\vec{C}$ is only approximately 0.2. This result indicates that despite that $\vec{C}$ may not be fully correct, the distance between the copy numbers of the tumor subclones is more important to reconstruct the subclonal evolution tree than the actual copy numbers.

3.2 Including information across all samples becomes useful as the noise levels increase

Previously, we made the assumption that including information across multiple samples increases the accuracy with which $\vec{C}$ is inferred. The method includes this information by selecting the best copy number from an initial choice of 2 copy numbers per segment, which are selected per sample individually. The copy numbers that then maximize $\mathbb{P}\left(\vec{C}, \vec{\mu} | \overrightarrow{LAF}\right)$ when $\mathbb{P}(\vec{C})$ is computed across all samples are selected as the final choice. It is interesting to investigate if allowing the model to select the best copy number from 2 initial choices is useful at all (semi-greedy), or if the best copy number with the highest probability per sample individually (greedy) is sufficient. The ratio with which the first copy number is selected compared to the second is shown in Fig. 8A. The ratio at which the second copy number is selected compared to the first is shown in Fig. 8B. From these figures, we observe that the first copy number with the highest probability is selected in approximately 99% of all segments in a sample at noise levels of 0 and 0.02. Starting from a noise level of 0.04 standard deviations, the copy number with the second highest probability is chosen more often (Fig. 8B). This result indicates that including information across samples becomes more useful as the noise levels increase. As the second best copy number is chosen more often for noise levels higher than 0.04 standard deviations, it may be useful to let the model choose from 3, or even more copy numbers to increase reconstruction accuracy with noise levels higher than 0.02 standard deviations. However, as noise levels above standard deviations of 0.02 are uncommon in our real dataset, we chose to not extend the number of copies to choose from per segment beyond 2.

We furthermore present the results of inferring $\vec{C}$ and $\vec{A}$ with the greedy approach in Section B.5 of the Supplementary data.

3.3 Real data: Testicular germ cell tumors

We applied the method to a case of 2 chemotherapy-resistant testicular germ cell tumors. No ground truth is known for the tumors. Furthermore, the absence of a method for inferring subclonal $\vec{C}$ and $\vec{\mu}$ from datasets without copy number information complicates comparisons to existing methodology. Therefore, our interpretations rely on existing knowledge

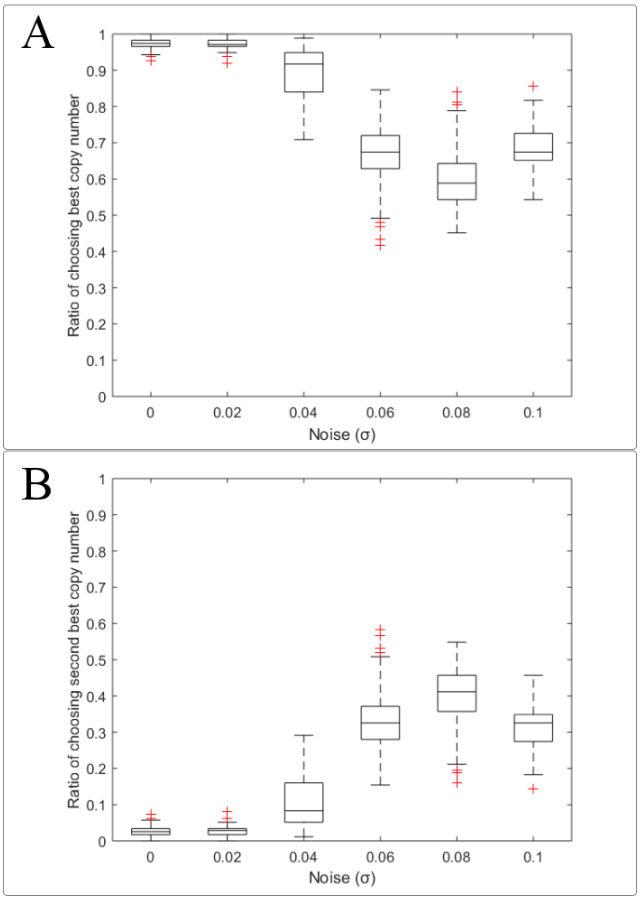

Fig 8. Ratio of how often (A) the best copy number is selected for the final $\vec{C}$ and (B) how often the second best copy number is selected for the final $\vec{C}$.

about the development of the tumors. A more detailed analysis of the genomic events identified in these tumors is provided in Section C.2 of the Supplementary data. For both cases, the reconstructed subclonal evolution trees shown here were reconstructed using the event distance based on $\vec{A}$ and the somatic variants. From expert reviews and comparisons to existing knowledge, these trees were more consistent with our expectations of the tumor development than the trees reconstructed using other combinations of $\vec{C}$, $\vec{A}$ and somatic variants. The results for the other combinations that can be used to reconstruct subclonal evolution trees can be found in Section C.3 of the Supplementary Data.

Expectations

As was discussed in the Introduction and shown in Fig. 2B, expectations about the development of testicular germ cell tumors are well-described in literature. First of all, we expect that samples of the precursor, carcinoma in situ, will genetically be most similar to the normal sample (peripheral blood), as the precursor is not expected to have gained as many somatic mutations as subtypes of the tumor that are formed at later stages of development. This carcinoma in situ precursor is assumed to develop into embryonal carcinoma, which may be able to differentiate into teratoma, yolk sac tumor or choriocarcinoma. In neither of the cases discussed below a sample of choriocarcinoma is present. We expect to observe a similar pattern in the reconstructed subclonal evolution trees, where carcinoma in situ is the parent of embryonal carcinoma, and embryonal carcinoma can form teratoma or yolk sac tumor. As somatic mutations and copy number

variations are accumulated over time as shown in Fig. 2B, we also expect that the later subtypes, such as teratoma and yolk sac tumor will be the furthest away from the normal sample in the tree. Furthermore, both cases contain samples of the macrodissected primary tumor (nonseminoma), which is a mixture of multiple subtypes of the tumor. As the measurements are therefore also an average of the genomes of the subtypes that are present in the mixture, we expect the genome to be more complex and therefore the furthest down in the tree. However, the exact placement in the tree could give insight into which subtypes that have been sampled are represented the most in the mixture.

Case 1: T6107

For the first case, we reconstructed a subclonal evolution tree using a total of 15 samples of the tumor. As can be seen in Fig. 7A**,** the reconstructed tree resembles the expectations discussed above. We see that the peripheral blood samples are at the root of the tree, corresponding with the assumptions that tumor development starts from normal, healthy cells. The two peripheral blood samples are highly similar as these are both composed of healthy cells, and therefore the choice of placing the first sample at the root of the tree rather than the second is arbitrary. We see that the carcinoma in situ and floating carcinoma in situ samples are indeed connected to the peripheral blood samples as expected. In our expectations we defined that the carcinoma in situ sample should be placed as the parent of the embryonal carcinoma samples, as carcinoma in situ is expected to be the precursor that is reprogrammed to form embryonal carcinoma. However, these relations are not immediately present in the tree: instead, we see that the second and fourth embryonal carcinoma samples are defined to be children of the peripheral blood sample rather than carcinoma in situ. This behavior can potentially be explained by the lack of ancestral nodes in the tree, which represent early precursor subclones in the tumor that have not been sampled, or may not even be present in the tumor bulk anymore at time of sampling. In theory, there could actually exist an additional step in the tumor development between the peripheral blood sample and the second and fourth embryonal carcinoma samples, which is also the parent of the two carcinoma in situ samples. In this proposed explanation, we assume that since carcinoma in situ is a precursor, and samples have been acquired after these precursors initially formed in the tumor, additional mutations may have accumulated over time in the carcinoma in situ that set them apart from the ancestral node that is common to the carcinoma in situ and second and fourth embryonal carcinoma samples.

Additional evidence for the existence of this ancestral node is provided by the positioning of the metastasis sample. This sample contains only 1 somatic variant that is also found in the carcinoma in situ samples (see Supplementary Figure C3A). Therefore, it is reasoned that the metastasis, which was found 15 months after removal of the primary tumor, has likely originated from an early precursor lesion that remained after treatment, but that has never been sampled. This again implies a possible hidden precursor subclone that follows immediately after the peripheral blood sample.

According to the reconstructed tree, the yolk sac tumor and teratoma subtypes originated from an embryonal carcinoma, which matches our initial expectations of the tumor development.

The fourth embryonal carcinoma sample is found in a different branch than the other embryonal carcinoma samples, which can be explained by the lack of one somatic variant in the fourth sample compared to the other samples. The infinite sites assumption, which states that somatic variants are not lost once gained, complicates the positioning of this sample in the tree. Due to the lack of abundant measurements around somatic variants, we are unable to accurately determine if this somatic variant has never been present, or if it has been lost together with the chromosomal region it was located on. Therefore, the placement of this sample ought to be interpreted with caution.

Finally, we see that the samples of the primary tumor are indeed placed at the bottom of the tree, and that the samples genetically mostly represent embryonal carcinoma. As the measurements in the primary tumor are averaged across all the subtypes that are present in the sample, it is difficult to tell if the average of the measurements makes the tumor similar to embryonal carcinoma, or if the sample actually mostly consists of embryonal carcinoma components.

Case 2: T3209

For the second case, we reconstructed the subclonal evolution tree from a total of 25 samples (Fig. 7B). As for the first case, we do not observe that the carcinoma in situ samples are inferred to be the parent of embryonal carcinoma samples as was initially expected. The tree indicates that the fourth teratoma sample is the most likely precursor of many other sampled components. This does not follow the expected model where embryonal carcinoma is the first precursor after carcinoma in situ, where embryonal carcinoma can then differentiate into other subtypes including teratoma. If we examine the inferred $\vec{C}$ and $\vec{A}$ for this sample (not shown here), we observe that most inferred copy numbers are 2, or close to 2, with alleles AB. Therefore, the distance of this teratoma sample to the peripheral blood is small. Furthermore, this teratoma sample has an equal number of somatic variants as all other samples but the peripheral blood, carcinoma in situ and fourth embryoid body samples. If the fourth teratoma sample lacked somatic variants compared to other samples, the assumption that somatic variants are never lost would have been useful to place the teratoma sample at the start of the tree. However, as there are no differences in the presence of somatic variants compared to the previously mentioned samples, the somatic variants are not useful in predicting the expected placement of the fourth teratoma sample in the tree. Furthermore, the actual truth about where the sample should be positioned in the tree is unknown. As teratoma components consist of somatic tissue, the inferred $\vec{C}$ could in fact be mostly correct and indicate a reversal back to a genome that is similar to the normal genome. Furthermore, as somatic tissue is difficult to distinguish from normal tissue, the estimated possible normal cell contamination of 27% could in fact be much higher than predicted by the model.

Most embryoid body samples cluster together in the tree, with the exception of the fourth and seventh embryoid body sample. The fourth embryoid sample lacks a somatic variant compared to the other embryoid body samples and has an additional loss of heterozygosity region on chromosomes 1 and 5 (segment 5A, see Supplementary Figure C3B). These characteristics sets the genome of this component apart from the other sampled embryoid body components. However, as the lesser allele frequencies of the regions other than the aforementioned and the rest of the somatic variants are highly similar between the embryoid body components, it is more likely that all embryoid body components in the tumor share the same precursor and that the fourth embryoid body sample should actually in the tree be connected to another embryoid body sample.

The seventh embryoid body sample has lesser allele frequency measurement patterns similar to the other embryoid body samples (see Supplementary Figure C3B), but overall seems to have more regions that have lesser allele frequencies of approximately 1/2. Therefore, the model infers that many segments with a median lesser allele frequency of 1/2 have a copy number of 2, again making the component seem more like a normal, healthy cell. Whether the inferred copy numbers and alleles are actually true or not is unknown.

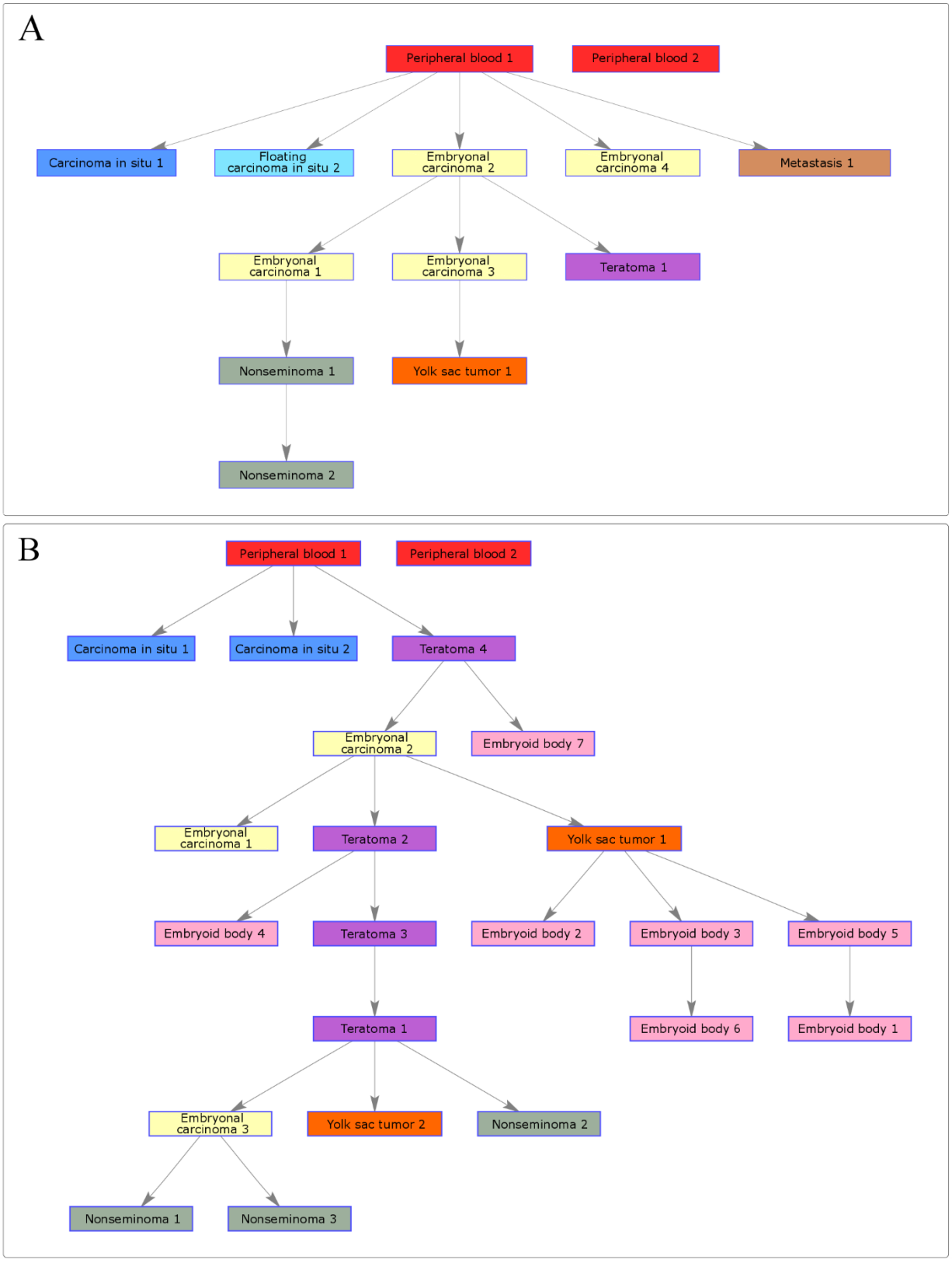

Fig. 9. Clonal evolution trees for the two testicular germ cell tumor cases, using the event distance based on the alleles and including somatic variants as weights. The same subtypes are represented with the same color. (A) Clonal evolution tree for case T6107. **(B)** Clonal evolution tree for case T3209.

4 Discussion

In this article, we have described TargetClone, a method developed to infer the most likely copy numbers ($\vec{C}$), frequency ($\vec{\mu}$) and alleles ($\vec{A}$) of subclones in a sample from targeted sequencing data. First, we demonstrated that in our simulation datasets, the method can infer $\vec{C}$ with a median accuracy of approximately 0.9 for noise levels of at most 0.02 standard deviations, which are typical noise levels for high quality frozen samples. $\vec{A}$ can be inferred with a median accuracy of approximately 0.9 when no noise is present in the measurements, while the median reconstruction accuracy decreases to approximately 0.7 as the noise level increases to 0.02 standard deviations. Furthermore, the median accuracy of inferring $\vec{\mu}$ is never lower than 0.7 and approximates 0.95 and 0.9 at noise levels of 0 and 0.02 standard deviations, respectively. Additionally, we were able to reconstruct subclonal evolution trees with a median accuracy of 0.8 with distance matrices based on a combination of the $\vec{C}$ that is inferred by the method and a binary representation of the presence and absence of somatic variants that are measured in the samples.

Second, we showed that $\vec{C}$ and $\vec{A}$ can be more accurately reconstructed at noise levels above 0.04 when we first keep the two best $\vec{C}$ per sample, and finally select the best $\vec{C}$ for which the overall distance based on the copy numbers is minimized between all samples.

Third, we reconstructed subclonal evolution trees for 2 testicular germ cell tumors with intrinsic resistance to chemotherapy. By comparing the resulting trees to descriptions of the development of testicular germ cell tumors in literature, we demonstrated that the trees were mostly consistent with our expectations.

Despite these promising results, the method has a number of important limitations. Most importantly, the median accuracy with which $\vec{C}$ and $\vec{A}$ are inferred decreases to approximately 0.18 and 0.08 at a noise level of 0.1, respectively. Therefore, we recommend to exclude any samples with high noise levels (which is typical in for example formalin-fixed paraffin-embedded samples) from analysis with TargetClone.

In addition, the median reconstruction accuracy of $\vec{C}$ and $\vec{A}$ decreases in the presence of copy number ambiguities. When the lesser allele frequency measurements are not affected by noise, the ambiguity score was no higher than 0.95 on average. Thus, our method is unable to resolve copy number ambiguities for approximately 5% of all segments per sample. Therefore, it is important to be careful when interpreting the inferred $\vec{C}$, $\vec{A}$ and $\vec{\mu}$ when the measured lesser allele frequencies in a sample could represent multiple possible underlying copy numbers in the tumor subclone.

Furthermore, we make the assumption that all samples consist of only one tumor subclone and is potentially contaminated with normal cells. However, even when samples have been microdissected, it may initially not always be known if the samples consist of only one tumor subclone or if multiple tumor subclones could be present. The current model of TargetClone does not support a good framework that can handle the inference of multiple subclones in a sample. When inferring the $\vec{C}$ and $\vec{\mu}$ that maximize the probability of observing the measured lesser allele frequencies in a sample, we perform an exhaustive search across all possible $\vec{\mu}$. If the sample is a mixture of normal cells and one tumor subclone, the model searches through 101 possible $\vec{\mu}$. If samples are a mixture with two or three tumor subclones, the model will search through 5151 and 17685 possible $\vec{\mu}$, respectively. The number of $\vec{\mu}$ to search through thus becomes too large, and an exhaustive search will require too much computational time. Ideally, it would still be possible to apply TargetClone even when it is not possible to assume that a sample contains only one tumor subclone. Furthermore, removing this assumption would make the method more widely applicable to studies in which samples are not microdissected and thus may contain multiple subclones, yet saving costs as we have shown that subclonal evolution can be reconstructed from targeted sequencing data. To conclude this article, we present some ideas for future work that may improve the usability of TargetClone.

5 Future work

As discussed in the previous section, the main bottleneck of TargetClone is the limitation to one subclone per sample, which is introduced by the exhaustive search through all possible $\vec{C}$ and $\vec{\mu}$ for each sample. An important future improvement may be to explore the benefits of inferring $\vec{C}$ and $\vec{\mu}$ with the use of optimization algorithms. Optimization algorithms will allow us to more efficiently search through all possible combinations of $\vec{C}$ and $\vec{\mu}$ without the need for an exhaustive search. One example of such algorithms is Expectation Maximization (EM), which has already been applied by the authors of Clomial to a similar problem of identifying tumor subclones and their frequencies in heterogeneous samples from variant allele frequencies measured for somatic variants[9]. The biggest difference of Clomial compared to our model is that Clomial is restricted to inferring a binary $\vec{C}$ matrix that indicates if a somatic variant is present or not in a tumor subclone. As our $\vec{C}$ matrix may contain any copy number *k* between a predefined *kmin* and *kmax,* the complexity of our model is higher than for Clomial, and may introduce additional challenges on the computational level. Nevertheless, optimization algorithms may be useful in solving other limitations of the method.

The first other limitation is that we make the assumption that the lesser allele frequency measurements have been assigned to segments that correspond to a region with a copy number and combination of alleles that are different from the adjacent segments. However, especially when copy number information is lacking for the samples and the lesser allele frequency measurements are affected by sequencing noise, it is not always clear what the exact locations are where segments begin and end. This factor makes it difficult to always provide a correct segmentation. As a result, our method will infer a copy number and alleles for a segment that may actually consist of two regions with different copy numbers and alleles. Especially if there is a rather large difference between the lesser allele frequency measurements between the two regions, say approximately 0.33 and 0.5, for example the mean of these measurements will be around 0.4. Thus, the inferred copy number and alleles for the segment will also be an average across the two regions and does not accurately represent the underlying genomic constitution. However, a more efficient algorithm may allow us to automatically update the segmentation in regions where it is likely that the initial segmentation is incorrect.

Second, the model assumes that all segments are independent and infers the best copy number and alleles per segment individually. The authors of MEDICC demonstrated that dealing with dependencies between segments can improve the reconstruction of copy number profiles in tumor samples[26]. For example, when two adjacent segments in one sample have a copy number of 1 and 2, respectively, and the same segments have a respective copy number of 0 and 1 in another sample, it may very well be that the two segments were affected by a loss at once. Our method can currently not work with dependencies between segments as this would mean that for each segment, at least the copy numbers of the two adjacent segments need to be inferred as well at the same time. Thus, per $\vec{\mu}$ and per segment, the model would already need to compute the likelihood for 7^3 $\vec{C}_i$. Therefore, if it would require less computational time to search through these

combinations with the use of an optimization algorithm, it may be possible to include dependency into the model and potentially improve the accuracy with which $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ are inferred.

Finally, TargetClone now reports only the $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ for which the likelihood of the lesser allele frequencies of a sample are maximized and reconstructs subclonal evolution trees based on the best solutions. However, there may exist other $\vec{C}$, $\vec{\mu}$ and $\vec{A}$ that have lower probabilities, but may reveal additional information that is interesting for researchers to investigate. If it would be less time consuming to search through the entire landscape of $\vec{C}$ and $\vec{\mu}$, it will be easier to report multiple solutions as these are typically explored and scored by optimization algorithms before the solution with the global maximum is obtained.

In addition, we demonstrated that reconstructing subclonal evolution trees is not perfect for our simulation data. The main reason behind the low accuracies observed here is likely that we use distance matrices to reconstruct the trees. If $\vec{C}$ or $\vec{A}$ is incorrect on one segment, and the distance on one segment is inferred to be infinite from subclone 1 to subclone 2, then subclone 1 can never be the parent of subclone 2. To overcome sensitivity to errors on segments, it may be useful to explore different methods to reconstruct subclonal evolution trees. For example, it may be useful to start with a tree where subclones are positioned according to their similarity, where branches are introduced only as soon as a relationship between subclones is not possible.

We here presented TargetClone, a method that can infer the copy numbers, alleles and frequency of subclones in multiple tumor samples and reconstruct subclonal evolution trees. We demonstrated that despite the existence of a number of discussed limitations, there is definitely potential in reconstructing the subclonal evolution of tumors from targeted sequencing data.

References

1. El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **31,** i62–i70 (2015).
2. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Rep.* **7,** 1740–1752 (2014).
3. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
4. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007 (2012).
5. Miller, C. a. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput. Biol.* **10,** e1003665 (2014).
6. Deshwar, A. G. *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16,** 35 (2015).
7. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
8. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
9. Zare, H. *et al.* Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLoS Comput. Biol.* **10,** e1003703 (2014).
10. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30,** 50–60 (2014).
11. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24,** 1881–93 (2014).
12. Purdom, E. *et al.* Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29,** 3113–3120 (2013).
13. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14,** 618–630 (2013).
14. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer Evolution: Mathematical Models and Computational Inference. *Syst. Biol.* **64,** e1–e25 (2015).
15. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15,** 35 (2014).
16. Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.* **41,** e165–e165 (2013).
17. Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* **30,** i78–86 (2014).
18. Larson, N. B. & Fridley, B. L. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29,** 1888–1889 (2013).
19. Oesper, L., Mahmoody, A. & Raphael, B. J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **7821 LNBI,** 171–172 (2013).
20. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11,** 396–398 (2014).
21. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274,** 998–1001 (1996).
22. Espina, V., Heiby, M., Pierobon, M. & Liotta, L. a. Laser capture microdissection technology. *Expert Rev. Mol. Diagn.* **7,** 647–57 (2007).
23. Grada, A. & Weinbrecht, K. Next-Generation Sequencing: Methodology and Application. *J. Invest. Dermatol.* **133,** 1–4 (2013).
24. Peng, Q., Vijaya Satya, R., Lewis, M., Randad, P. & Wang, Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics* **16,** 589 (2015).
25. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14,** R51 (2013).
26. Schwarz, R. F. *et al.* Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Comput. Biol.* **10,** e1003535 (2014).
27. Looijenga, L. H. Human testicular (non)seminomatous germ cell tumours: the clinical implications of recent pathobiological insights. *J. Pathol.* **218,** 146–162 (2009).
28. Rijlaarsdam, M. a & Looijenga, L. H. J. An oncofetal and developmental perspective on testicular germ cell cancer. *Semin. Cancer Biol.* **29,** 1–16 (2014).
29. Sheikine, Y. *et al.* Molecular genetics of testicular germ cell tumors. *Am. J. Cancer Res.* **2,** 153–67 (2012).
30. Boublikova, L., Buchler, T., Stary, J., Abrahamova, J. & Trka, J. Molecular biology of testicular germ cell tumors: Unique features

awaiting clinical application. *Crit. Rev. Oncol. Hematol.* **89,** 366–385 (2014).

31. Rijlaarsdam, M. a. *et al.* Genome Wide DNA Methylation Profiles Provide Clues to the Origin and Pathogenesis of Germ Cell Tumors. *PLoS One* **10,** e0122146 (2015).
32. Oosterhuis, J. W. & Looijenga, L. H. J. Testicular germ-cell tumours in a broader perspective. *Nat. Rev. Cancer* **5,** 210–222 (2005).
33. Bartkova, J., Rajpert-De Meyts, E., Skakkebaek, N. E., Lukas, J. & Bartek, J. DNA damage response in human testes and testicular germ cell tumours: biology and implications for therapy. *Int. J. Androl.* **30,** 282–91; discussion 291 (2007).
34. Edmonds, J. Optimum branchings. *J. Res. Natl. Bur. Stand. Sect. B Math. Math. Phys.* **71B,** 233 (1967).