Simultaneous estimation of per cell division mutation rate and turnover rate from bulk tumour sequence data

Gergely Tibély[1,2], Dominik Schrempf[2], Imre Derényi[2,3], Gergely J. Szöllősi[1,2,4]

[1]MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary
[2]Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary
[3]MTA-ELTE Statistical and Biological Physics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary
[4]Institute of Evolution, Centre for Ecological Research, Konkoly-Thege M. út 29-33. H-1121 Budapest, Hungary

February 12, 2021

Abstract

Tumors often harbor orders of magnitude more mutations than heal thy tissues. The increased number of mutations may be due to an elevated mutation rate or frequent cell death and correspondingly rapid cell turnover leading to an increased number of cell divisions and more mutations, or some combination of both these mechanisms. It is difficult to disentangle the two based on widely available bulk sequencing data where mutations from individual cells are intermixed. As a result, the cell linage tree of the tumor cannot be resolved. Here we present a method that can simultaneously estimate the cell turnover rate and the rate of mutations from bulk sequencing data by averaging over ensembles of cell lineage trees parameterized by cell turnover rate. Our method works by simulating tumor growth and matching the observed data to these simulations by choosing the best fitting set of parameters according to an explicit likelihood-based model. Applying it to a real tumor sample, we find that both the mutation rate and the intensity of death is high.

Author Summary

Tumors frequently harbor an elevated number of mutations, compared to healthy tissue. These extra mutations may be generated either by an increased mutation rate or the presence of cell death resulting in increased cellular turn over and additional cell divisions for tumor growth. Separating the effects of these two factors is a nontrivial problem. Here we present a method which can simultaneously estimate cell turnover rate and genomic mutation rate from bulk sequencing data. Our method is based on the maximum likelihood estimation of the parameters of a generative model of tumor growth and mutations. Applying our method to a

human hepatocellular carcinoma sample reveals an elevated per cell division mutation rate and high cell turnover.

1 Introduction

Cancer is an evolutionary phenomenon within a host organism that unfolds on the timescale of years or more. New mutations can appear with each cell division, while cells can also die for reasons such as lack of nutrients or immune reactions. Due to the limitations of bulk sequencing, which only essays mutation frequencies for a population of cells from each tumor sample and does not resolve individual cells' genotype, basic evolutionary parameters. In particular, the cell turnover rate and per cell division mutation rate remain unknown, with estimated values spanning several orders of magnitudes [1].

While tumors can contain a large number of mutations, it is not clear whether this is due to an elevated mutation rate or frequent cell death, as frequent cell death results in more cell divisions, which, in turn, gives rise to more mutations. There are arguments for both cases [2, 3, 4, 5, 6], but distinguishing between these two alternatives is difficult becasue we cannot resolve the tumor's cell lineage tree from bulk sequencing data.

In previous work [2, 1], an elevated number of mutations was observed, but only the combined effect of the mutation rate and the death rate could be estimated. Williams et al. [7] targeted the problem of separating these two quantities by separately sequencing in bulk multiple samples from the same tumor thus resolving a coarse grained cell linage tree. However, it is not clear whether this approach resolves the cell lineage tree in sufficient detail to identify the regime of frequent cell death when the number of mutations is orders of magnitude larger than under growth without cell death.

Here, we describe a method to simultaneously estimate the per cell division mutation rate and the turnover rate (the ratio of death and birth rates) of a tumor from bulk sequencing data. The estimation is based on a maximum likelihood fit of the parameters of a birth-death model to the measured mutant and wild-type read counts. While requiring only a single tumor-normal sample pair, the fitting procedure can differentiate between death rates, which are extremely close to the critical value where the birth rate equals the death rate, resulting in accurate estimation of the mutation rate across orders of magnitudes. The rest of the paper is structured as follows. After introducing our model and the fitting procedure in Sec. Methods, we assess model accuracy on simulated data in Sec. Results on synthetic data. Results on empirical data are described in Sec. Results on empirical data, and conclusions are given in Sec. Discussion.

2 Methods

Model

We describe the evolution of tumor cells with the cell linage tree, i.e. the bifurcating tree traced out by cell divisions. As cells that have died cannot be observed by sequencing we consider the tree spanned by surviving cells. The leaves of this tree correspond to extant cells and internal nodes to observed cell

divisions. To model the descendance of the extant cells, we employ the conditioned birth-death process with birth rate α and death rate β and a fixed number of cells n sampled [8]. We measure branch lengths in numbers of expected birth events (the product of birth rate and time). Consequently, the role of the birth rate α can be considered as a scaling constant that sets the unit of time and we consider it to be equal to 1 without loss of generality. As result the death rate determines to the turnover rate: $t = \beta/\alpha = \beta$.

Mutations occur with a rate μ per site per cell division, mutations are considered neutral and we neglect the probability that a site is hit by a mutation by more than one time, in accordance with the infinite site hypothesis.

The data available from bulk sequencing is the mutant and wildtype readcounts of sites. Therefore, we will use the site frequency spectrum, which can be estimated from readcount data, to separate the effects of the mutation rate and cell death intensity. The site frequency spectrum reflects the branch length distribution of the tumors cell lineage tree. The tree's leaves are the sequenced cells, and its root is the most recent common ancestor of these cells. Changing the turnover rate modifies the shape of the cell linage tree by changing the relative lengths of branches closer to the root compared to terminal ones and as a result modifying the site frequency spectrum. Changing the mutation rate, on the other hand, simply results in more mutations, thus leaving the shape of the tree, and by extension the overall shape of the site frequency spectrum unchanged (see Fig. 1). It should be noted, however, that the information we will use is more detailed than the site frequency spectrum, namely, the read count pairs of mutated sites, which contain more information than just one rational number. E.g., 10 mutant reads out of 30 reads and 1 out of 3 both contribute 1 mutation to the frequency 1/3, while the uncertainty of the first case is significantly lower than for the second case. It also makes quite straightforward to include nucleotide-dependent transition probabilites, or trinucleotide context-based effects.

Site frequency spectra derived from tumors alos contain the effects of the ploidy of the sites and the contamination of the sample by normal cells. The corresponding spectrum is termed Variant Allele Frequency (VAF) spectrum. VAF frequencies are also affected by the finite sequencing depth, which gives rise to a stochastic variation in the observed allele frequencies.

Throughout the paper, the following notation is used: Branches of the cell lineage tree are denoted by the index k, the length of branch k is denoted by l_k and $L = \sum_k l_k$ denotes the sum of all branch lengths in the tree. The numbers of mutations per site from cell divisions along branch k are Poisson distributed and their sum is also Poisson distributed. The number of expected cell divisions along branch k is l_k, therefore, the distribution of the number of mutations per site on branch k is a Poisson distribution with parameter $l_k\mu$. Similarly, the total number of mutations is Poisson distributed with parameter $L\mu n_{\text{sites}}$.

Inference

To compare different combinations of mutation and turnover rates describing the observed empirical data we employ a maximum likelihood approach. First, we derive the likelihood of the observed data, $\mathcal{L}(D|\mu, t)$, as a function of the mutation rate and the turnover rate. As described below we maximize this likelihood function averaged over a random sample of cell lineage trees with

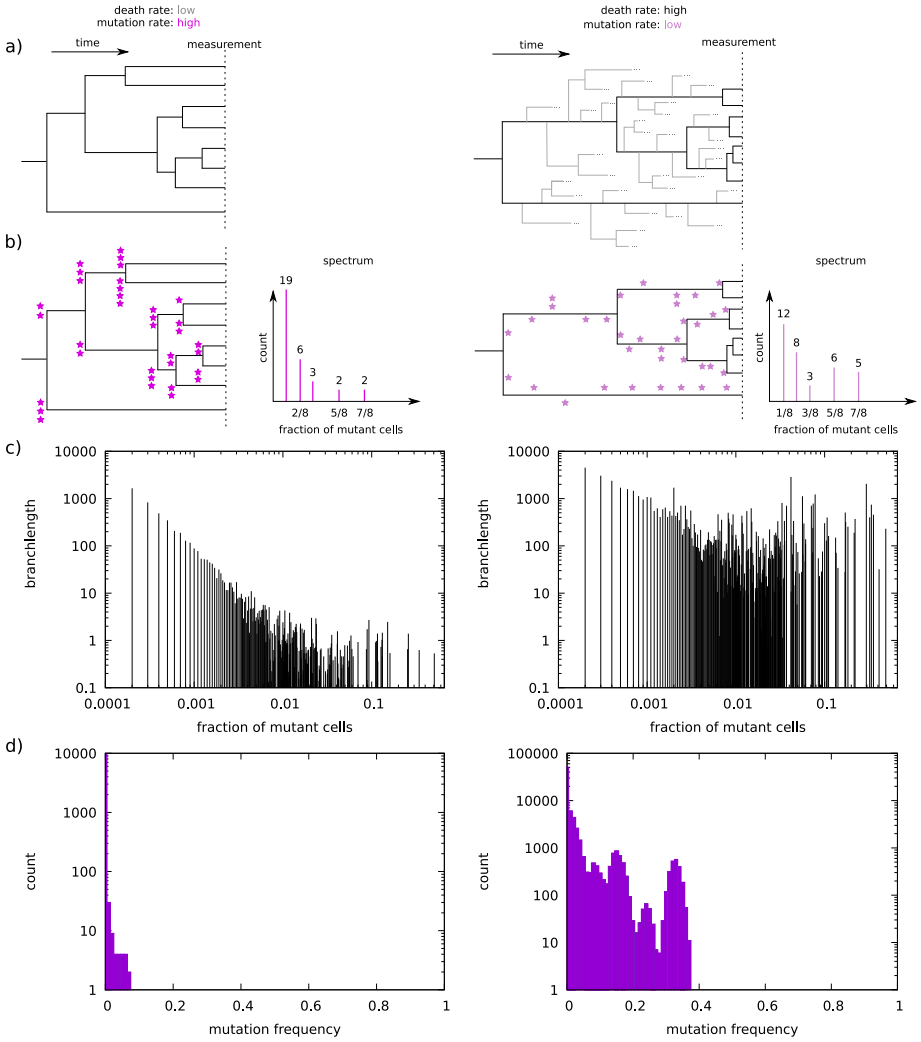

Figure 1: Two possible scenarios for the generation of mutations along cell lineage trees. **a)**: different turnover rates lead to different lineage tree shapes. Bifurcations are cell divisions, leaves are cells comprising the bulk sequencing sample. Note that the (surviving) tree topologies are the same, only branch lengths differ. **b)**: mutations, symbolized by purple stars, accumulate at cell divisions. High turnover rate and low mutation rate can lead to the same number of observed mutations as low turnover rate and high mutation rate, however, the mutation spectrum of the trees are different. **c)**: For simulated trees of 10000 leaves, the differences in the branch length distribution are clearly visible. **d)**: VAFs of the mutation spectra. Fractions of mutant cells are binned (note the logarithmic scale). Ploidy is set to two, contamination is zero. Simulated sequencing depth is 1000.

fixed turnover rate t in order to estimate the parameters t and μ that are most likely to have generated the observed data.

First we derive $\mathcal{L}(D|\mu, \mathcal{T})$ the likelihood of the observed data for a fixed cell lineage tree $\mathcal{T}$. It is assumed that sites collect mutations independently of each other, consequently, $\mathcal{L}(D|\mu, \mathcal{T})$ takes the form of a product over sites:

$$\mathcal{L}(D|\mu, \mathcal{T}) = \prod_{\substack{i \text{ in} \\ \text{sites}}} p(m_i|\mu, \mathcal{T}, r_i) \tag{1}$$

where m_i is the number of reads exhibiting a mutation at site i, and r_i is the total number of reads covering site i. To calculate the probability of observing m_i mutant reads out of a total of r_i reads we consider the following alternatives: i) if $m_i = 0$ either a mutation occurred, with probability $P_{\text{mut}}(\mu, L) = 1 - \exp(-\mu L)$ (see also Sec. Methods), but no mutant read, i.e. $m_i = 0$ was observed out of r_i reads with probability $F[0, r_i, \mathcal{T}]$, or no mutation occurred with probability $1 - P_{\text{mut}}(\mu, L)$ or ii) a mutation occurred with probability $P_{\text{mut}}(\mu, L)$ and m_i mutant reads where observed out of r_i, with probability $F(m_i, r_i, \mathcal{T})$:

$$p(m_i|\mu, \mathcal{T}, r_i) = \\ = \begin{cases} P_{\text{mut}}(\mu, L) \cdot F(0, r_i, \mathcal{T}) + (1 - P_{\text{mut}}(\mu, L)), & m_i = 0 \\ P_{\text{mut}}(\mu, L) \cdot F(m_i, r_i, \mathcal{T}) & m_i > 0 \end{cases} \tag{2}$$

To compute the probability $F(m, r, \mathcal{T})$ of observing m mutant reads out of r total reads given the cell linage tree $\mathcal{T}$, we assume that the mutant reads descend from a single mutation that occurred at somepoint along branch k, which has a length l_k and from which a fraction f_k of sequenced cells descend, and take the sum over all branches:

$$F(m, r, \mathcal{T}) = \sum_k \frac{l_k}{L} \cdot \text{Binom}(m, r, f_k) = \sum_k \frac{l_k}{L} \cdot \binom{m}{r} \cdot f_k^m (1 - f_k)^{r-m}, \tag{3}$$

where $L = \sum_k l_k$ and $\text{Binom}(m, r, f_k)$ is the probability mass function of the binomial distribution, i.e. the probability of getting exactly m successes in r independent Bernoulli trials with a probability of success f_k. We consider multiple mutations at the same site as a single mutation, and neglect all subsequent mutations after the first one. In all our applications we verified that $\mu L \ll 1$ is fulfilled.

To take into consideration sequencing errors, we must consider that they can lead to an excess of false mutation reads. To account for sequencing error, we introduce a parameter ε denoting the probability of a sequencing error at each position of each read. For $\varepsilon > 0$ new possibilities arise: mutant reads can be real mutants or false mutants, due to sequencing errors. It is also possible that all mutant reads at a site are false mutants, and there may or may not be a real mutation. We neglect the case when two or more mutations happen at the same site. Each position in each read can now be either wild type, false wild type, real mutant or false mutant,

$$
\begin{aligned}
&p(m_i|\mu,\mathcal{T},r_i,\varepsilon) \approx \\
&\qquad \approx P_{\text{mut}}(\mu,L)\sum_k \frac{l_k}{L}\frac{r_i!}{(r_i-m_i)!m_i}\big(f_k(1-\varepsilon)+(1-f_k)\varepsilon\big)^{m_i}\cdot \\
&\cdot\big((1-f_k)(1-\varepsilon)+f_k\varepsilon\big)^{r_i-m_i}+(1-P_{\text{mut}}(\mu,L))\binom{r_i}{m_i}\varepsilon^{m_i}(1-\varepsilon)^{r_i-m_i} \qquad (4)
\end{aligned}
$$

Note that Eq. 4 contains the $m_i = 0$ case.

Finally, we introduce the ability to differentiate mutant types, to conform the case of real DNA, which has 3 possible mutant types. So far, it was assumed that each site can have 2 states, wild type or mutant, corresponding to a DNA consisting of only two types of nucleotides, instead of four. Therefore, instead of the mutant read count m, we introduce three mutant read counts, corresponding to the three possible mutant types, $m^{(1)}, m^{(2)}, m^{(3)}$. Consequently, the input data now consists of triplets of mutant read counts, instead of one scalar mutant readcount. This leads to the use of a multinomial distribution, with four states: wild type and 3 mutant types. The possibility of more than one real mutant types at the same site is still neglected, being very rare, technically a second-order process in the mutation probability of one site. We also neglect the probability of more than one error hitting the same site. The likelihood function at a single site is then

$$
\begin{aligned}
&p(m_i^{(1)},m_i^{(2)},m_i^{(3)}|\mu,\mathcal{T},r_i,\varepsilon) \approx \\
&\quad \approx P_{\text{mut}}(\mu,L)\sum_k \frac{l_k}{L}\frac{1}{3}\sum_{j=1}^{3}\text{Mult}\Bigg(\Big(m_i^{(j)},m_i^{(j+1)},m_i^{(j+2)},r_i-\sum_{j'}m_i^{(j')}\Big); \\
&\qquad r_i;\Big(p_{\text{M}}^{(j)}(f_k,\varepsilon),p_{\text{M}}^{(j+1)}(f_k,\varepsilon),p_{\text{M}}^{(j+2)}(f_k,\varepsilon),p_{\text{W}}(f_k,\varepsilon)\Big)\Bigg)+ \\
&\quad + (1-P_{\text{mut}}(\mu,L))\cdot\text{Mult}\Bigg(\Big(m_i^{(1)},m_i^{(2)},m_i^{(3)},r_i-\sum_j m_i^{(j)}\Big);r_i; \\
&\qquad \Big(p_{\text{M}}^{(j)}(0,\varepsilon),p_{\text{M}}^{(j+1)}(0,\varepsilon),p_{\text{M}}^{(j+2)}(0,\varepsilon),p_{\text{W}}(0,\varepsilon)\Big)\Bigg) \qquad (5)
\end{aligned}
$$

and

$$
p_{\text{M}}^{(j)}(f_k,\varepsilon) = f_k(1-\varepsilon)+(1-f_k)\varepsilon/3 \qquad (6)
$$

$$
p_{\text{M}}^{(j+1)}(f_k,\varepsilon) = f_k\varepsilon/3+(1-f_k)\varepsilon/3 \qquad (7)
$$

$$
p_{\text{M}}^{(j+2)}(f_k,\varepsilon) = f_k\varepsilon/3+(1-f_k)\varepsilon/3 \qquad (8)
$$

$$
p_{\text{W}}(f_k,\varepsilon) = f_k\varepsilon/3+(1-f_k)(1-\varepsilon) \qquad (9)
$$

where $(j+1)$ and $(j+2)$ denote the other two possible mutant types with cyclic notation $(j) = (j+3)$, and Mult is a multinomial distribution the arguments of which denote (random variables), n_{draws}, (event probabilities). The factor of $1/3$ is due to the assumption that only one true mutation can be present at one

site, and each of the 3 possible mutated forms has the probability 1/3. Each $\mathrm{Mult}(m^{(j)} \ldots)$ term is a conditional probability conditioned on the true mutant type being (j).

The straight-forward approach for treating the unknown cell linage tree as a nuisance parameter would be to average over all trees $\mathcal{T}$:

$$\mathcal{L}(D|\mu, t) = \sum_{\mathcal{T}} \mathcal{L}(D|\mu, \mathcal{T}) \cdot p_{\mathrm{BD}}(\mathcal{T}|t), \tag{10}$$

where $p_{\mathrm{BD}}(\mathcal{T}|t)$ is the probability of the cell lineage tree $\mathcal{T}$ given conditioned birth-death process with turnover rate t. Due to the very large number of possible trees the above average, however, is intractable and we must results to sampling a finite number of trees drawn from the conditioned birth-death process with fixed t. Based on empirical experience we found that using the geometrical mean of $\mathcal{L}(D|\mu, \mathcal{T})$ for a finte sample of trees sampled according to $p_{\mathrm{BD}}(\mathcal{T}|t)$ results in more robust inference. The geometric mean approximates the average probability of inference [29] or equivalently the average surprisal [30] of cell linage trees given the turnover rate t, which we denote

$$\bar{\mathcal{L}}(D|\mu, t) = \prod_{\mathcal{T}} \mathcal{L}(D|\mu, \mathcal{T})^{p_{\mathrm{BD}}(\mathcal{T}|t)} = \exp\left(\sum_{\mathcal{T}} p_{\mathrm{BD}}(\mathcal{T}|t) \ln \mathcal{L}(D|\mu, \mathcal{T})\right). \tag{11}$$

In practice, during inference of the turnover rate t and mutation rate μ the log-average over a finte number of trees drawn from the conditioned birth-death process with turnover rate t is maximized:

$$\ln \bar{\mathcal{L}}(D|\mu, t) = \frac{1}{n_{\mathrm{trees}}} \sum_{\mathcal{T}} \ln \mathcal{L}(D|\mu, \mathcal{T}) \tag{12}$$

Generating trees

To generate cell division trees from birth-death conditioned process, we use the ELynx software suite [11], which allows freely adjustable birth rates, death rates, and tree sizes.

Generating synthetic samples

For generating synthetic samples of read counts of mutated DNA sites, trees simulated by ELynx are used as genealogical trees of hypothetical tumors. For each site, first we determine the total readcount at that site. Then, we draw random numbers to check whether any of the branches contributes a mutation, according to the Poisson process described in Sec. Inference. If there is a mutation, the true mutant readcount is drawn from a hypergeometric distribution. The number of successes of the hypergeometric distribution is the number of leaves of the selected branch. The number of failures is the total number of leaves multiplied by ploidy and divided by the hypothetical purity of the sample, minus the number of successes. The number of trials is the total readcount. Finally, errors are introduced by drawing a quadruplet of readcounts ("wildtype errors") from a multinomial distribution, with probabilities $(\varepsilon/3, \varepsilon/3, \varepsilon/3, 1-\varepsilon)$, the number of trials is the wildtype readcount. "Mutant errors" are drawn from another multinomial distribution, with probabilites $(1-\varepsilon, \varepsilon/3, \varepsilon/3, \varepsilon/3)$, the

number of trials is the mutant readcount. The final readcounts are given by the sum of the two drawn quadruplets.

The mutation rate for different turnover rates is chosen such that the total number of observed real mutations should remain close to each other, i.e., the estimation algorithm should have a similar amount of input data.

Calculating the likelihood

The goal is to find the maximum of the likelihood as the function of the mutation rate, turnover rate, and the error rate, to be able to use it for estimating the mutation and turnover rates by Eq. 10. The input is the read counts of the DNA sites. We use pre-generated division trees from the ELynx suite at pre-determined turnover rate values. Between these pre-determined turnover rate values, the likelihood function is interpolated using cubic splines. In the case of synthetic input datasets, the tree used to generate the test dataset is never included in the likelihood calculation.

The maximum of the likelihood function for each fixed turnover rate value is obtained by optimizing the error rate using Brent's method, implemented in Julia's Optim package, and estimating the mutation rate from the number of input mutations and the branch lengths of the currently fitted tree. Only mutations having read counts high enough to exclude sequencing errors are taken into account in the mutation rate estimation. The estimated mutation rate is averaged over trees, using uniform weights. Therefore, the likelihood function is optimized for ε at different t values, which come from a pre-defined set, for which the trees can be generated in advance, avoiding the need for new trees at each step of the optimization process.

3 Results on synthetic data

No sequencing errors

Figs. 2 shows the estimated turnover rates and mutation rates as functions of the true turnover rates and mutation rates. The method can reasonably differentiate between datasets with different true turnover rates-mutation rates and estimate their values. Fig. 3 shows the joint estimation of mutation rate-turnover rate parameter pairs. The data points are arranged into lines, corresponding to constant numbers of observed mutations, obeying

$$N_{\text{obs mut}}(\mu, t) = \mu \cdot \mathbb{E}\left(\sum_k l(k) * \text{ccdf}(\text{Binom}(\overline{n}_{\text{seq}}, f_k), 0)\right)_{\text{trees}(t)} \tag{13}$$

where ccdf(...) is the complementary cumulative distribution function of a binomial distribution, evaluated at 0. The parameters of the binomial distribution are the average sequencing depth $\overline{n}_{\text{seq}}$ and the fraction of leaves under branch k, f_k. The expected value is taken over the trees generated with turnover rate t. The lines defined by Eq. 13 can be numerically approximated for any μ, t points, averaging over a number of trees.

We checked the dependency of the results on the sizes of the trees. According to Fig. 4, estimation of large true turnover rates becomes increasingly harder as the tree sizes decrease. This can be attributed to the fact that differences

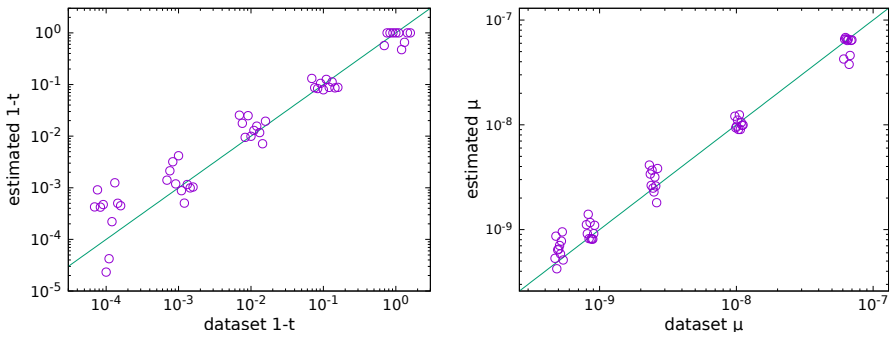

Figure 2: Estimated turnover rates (left) and mutation rates (right) for different true values. 10 synthetic datasets for each true value. The trees used for fitting have 10000 leaves, 10000 trees were used for each t value of the loglikelihood(t) measurements (see Fig. 7). The continuous line is a guide for the eye, corresponding to $y = x$. Points are slightly dispersed horizontally for clarity. Horizontal ordering of the data points is the same for both subplots, e.g., the rightmost point in each group of points corresponds to dataset no. 10 in both plots.

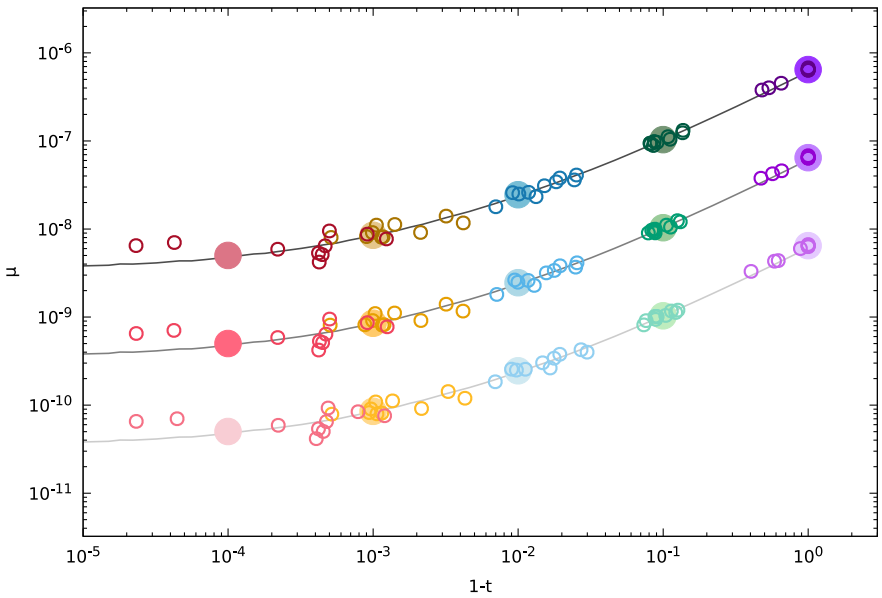

Figure 3: Joint estimations of mutation rate-turnover rate parameter pairs. 10 synthetic datasets for each true parameter pair, each of which is denoted by one color. True parameter values are indicated by large full circles. Solid lines show the numerical approximation of $\mu(1-t)$, for $N_{\text{obs mut}} = 5 \cdot 10^5, 5 \cdot 10^4, 5 \cdot 10^3$.

between the effects of very high turnover rates are observable on branches having a very small relative number of leaves, therefore large trees are required to

distinguish between high turnover rates.

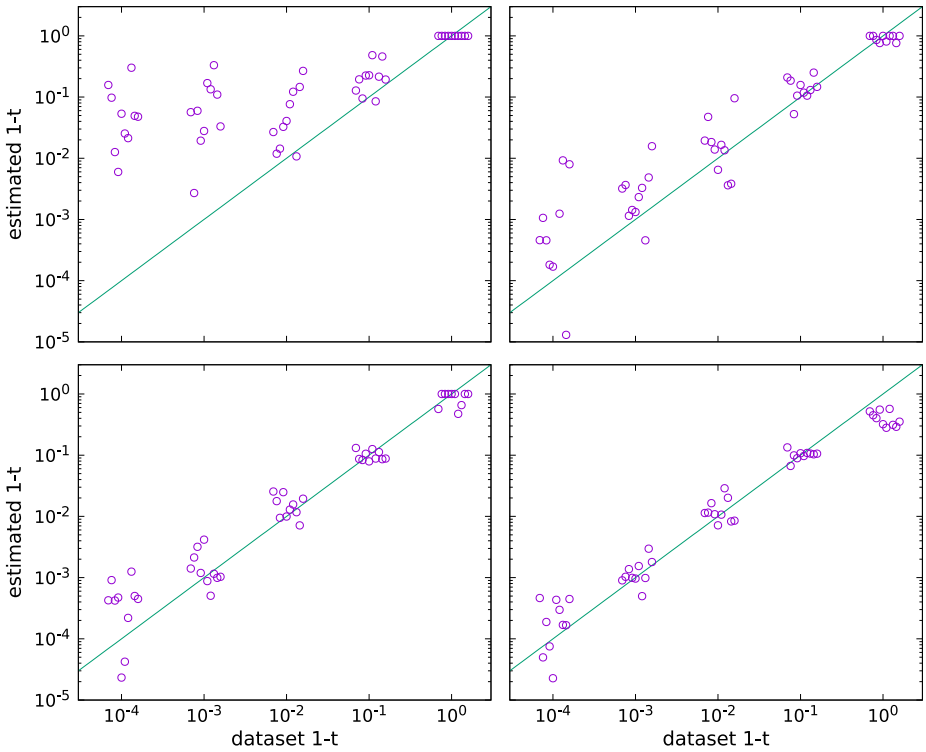

Figure 4: The effect of tree sizes on the estimations. Estimated turnover rates for different fitted trees sizes: 100 (**top left**), 1000 (**top right**), 10000 (**bottom left**), 100000 (**bottom right**).

The accuracies of the estimates for different datasets are not equal, besides the effect of the size of the trees in the case of high turnover rates. Differences between estimates for different datasets can be due to 3 possible factors: the trees used in the fitting process, the generated input data, or the tree used for generating the data. To check these factors, we chose a dataset which resulted in a turnover rate estimation deviating from the true value (dataset no. 7 for $1-t=1$ on Fig. 2, $1-t=0.47$). We calculated the estimated turnover rate values using 10 independent sets of 1000 trees. The estimations ranged from $1-t=0.39$ to 0.54. Therefore, the deviation of the estimate from 1.0 cannot be attributed to the sample of fitting trees. Then, we generated 10 more datasets using the same tree as for the original dataset. The estimated turnover rate values were between $1-t \in [0.44, 0.49]$, even more closely matching the original estimate. Consequently, the effect does not depend on the generated data but on the tree used to generate the data. It seems that the deviation of the estimates from the true turnover rates is due to the fluctuation of the shapes of the trees used for sample generation.

Effects of sequencing errors

To estimate the effect of sequencing errors, we calculated the estimations of the turnover rates, applying different amounts of errors to the same data (exactly the same mutant and wild type readcounts for each mutation). In this case, the error rate of the data was also estimated by the fitting procedure, along with the mutation and turnover rates. The influence of sequencing errors on the estimation of the turnover rates is shown on Fig. 5. For an error rate of 10^{-3}, which is frequently cited as the error rate of the Illumina sequencing technology [13], the estimated turnover rates can have significant deviations from the true values.

For lower error rates, the estimations approach the true values, however, outliers remain even for $\varepsilon = 10^{-8}$. We note that the loglikelihoods of the false estimations were better than those of the true values. We tried estimating the parameters while leaving out the least frequent mutations, to reduce the effect of errors, but the estimated parameters deviated significantly more from their true values.

4 Results on empirical data

To estimate the turnover and mutation rates of real tumors, a real human tumor sample is required. Due to the fitting method's sensitivity to high sequencing error rates, we need a sample which is sequenced using a very low error rate technology. Such samples are much less ubiquitous than those by the standard technology, and are usually restricted to very short genome segments, mostly nonhuman. Nevertheless, we found a sample of a human hepatocellular carcinoma (HCC) [14], which was sequenced using the o2n sequencing technology [15], providing error rates between 10^{-5}-10^{-8}, which is significantly lower than the 10^{-3} rate of the standard Illumina process. Besides the low error rate, the amount of sequenced positions is enough to cover the targeted region 3400x [15], which is also much better than those of standard quality datasets (typical sequencing depth is around 30). High sequencing depth results in more identified mutations and more precise mutation frequencies. Sequencing was targeted at a 410k basepair wide region of the genome, which is much narrower than the whole human genome. This is a typical shortcoming of low error rate sequencing methods. Still, as our fitting procedure is much more sensitive to the sequencing error rate than to the amount of the input mutations (compare the deviations on Figs. 3 and 5), the dataset provides significantly better input than typical sequencing data.

The raw sequencing data was preprocessed according to [15], using the code provided by the authors. The DNA contents of 10000 cells were sequenced [14], along with a sample of neighboring normal tissue. Mutations were called using VarScan 2 [16], which is flexible and easy to adapt to the requirements of the fitting procedure. 923383 sites remained after preprocessing, with sequencing depth being at least 8 (default for VarScan 2) in both tumor and normal samples. The distribution of sequencing depths is wide, ranging from 8 to 10853, with a mean of 904. For mutation calling by VarScan 2, the minimum number of mutant reads was set to 1 and the strand filter switched off. Although the number of false positives increases with these parameter choices, the resulting called mutations

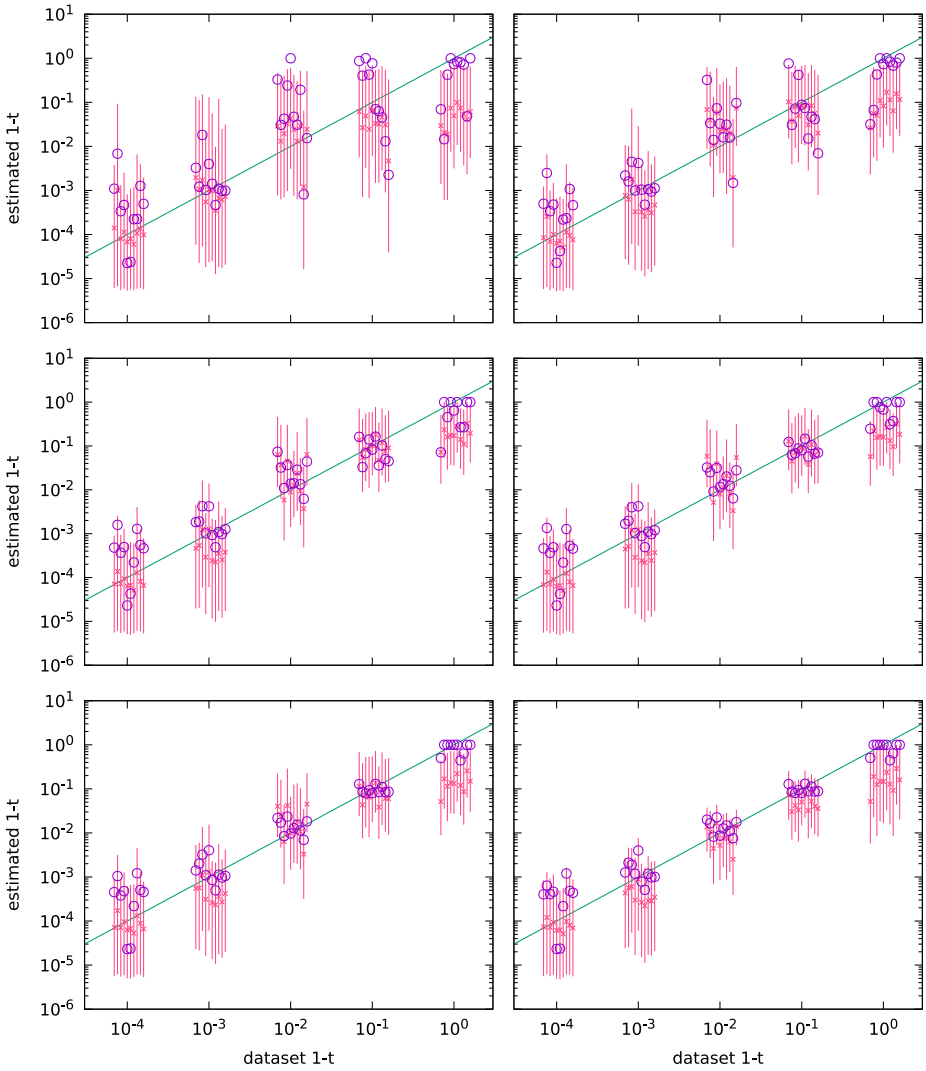

Figure 5: The effect of varying the error rate. Sequencing error rates are $\varepsilon = 10^{-3}$ (**top left**), 10^{-4} (**top right**), 10^{-5} (**middle left**), 10^{-6} (**middle right**), 10^{-7} (**bottom left**), 10^{-8} (**bottom right**). 10000 trees, tree size = 10000. X coordinates are slightly dispersed for clarity. Open circles are results corresponding to error rates used in the fit fixed to their true values, crosses correspond to error rates estimated by the parameter fit. Vertical lines show the ranges between the first and last 10-quantiles, based on 10000 beta value estimations. Each open circle-cross pair corresponding to the same dataset is vertically aligned.

correspond better to the error model of our fitting procedure than an error rate which changes sharply with threshold frequency or readcount values. The minimum variant frequency was set to 10^{-6} to include even the least frequent mutation. Purity was set to 0.85, in accordance with [14]. We also checked that the default somatic p-value threshold does not exclude any candidate somatic

mutations. Other parameter settings were unaltered from their default values. Mutation frequencies were corrected for copy number variation (CNV), using VarScan 2 with default parameters, and for ploidy of the sex chromosomes. CNV detection for targeted sequencing data is a more difficult task than for whole genome data, and VarScan 2 was found to be a stable performer [17]. Sites having multiple variant types (i.e., number of reads of wildtype plus most frequent mutant type being lower than the sequencing depth) were checked manually. Readcounts of all 4 possible genotypes were identified for all variant sites.

After all these steps, 2284 mutations were identified. The variant allele frequency spectrum is shown of Fig. 6.

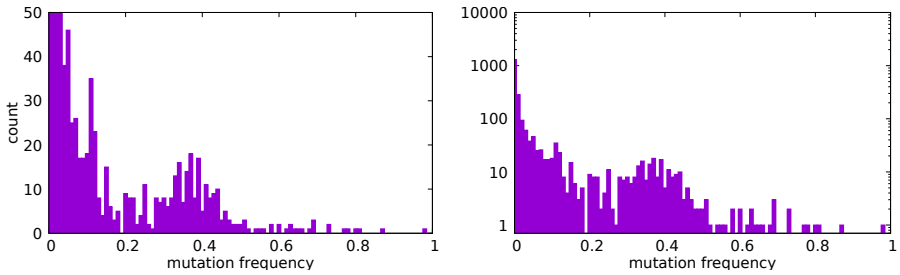

Figure 6: Variant allele frequency spectrum of a human hepatocellular carcinoma sample [14], obtained by o2n sequencing [15].

To estimate the sequencing error rate, the fitting procedure was applied to the mutation data with various fixed error rates in the range 10^{-8}-10^{-4}. The maximum of the loglikelihood corresponded to $\varepsilon = 10^{-7}$. It is a plausible value, as [15] estimated the error rate between 10^{-5}-10^{-8}. Based on estimating 10000 best error rates, the error rate should be between $9.3 \cdot 10^{-8}$ and $1.2 \cdot 10^{-7}$.

Having determined the error rate, we estimated the mutation and turnover rates, with the error rate fixed, using 10^5 trees. Fig. 7 shows the estimated turnover rate, $1 - t = 1.09 \cdot 10^{-3}$, and the mutation rate, $\mu = 9.26 \cdot 10^{-8}$ per site per cell division. Corresponding to the range of the error rate, the turnover rate ranges within $1.09 \cdot 10^{-3}$-$1.10 \cdot 10^{-3}$. The mutation rate ranges between $9.23 \cdot 10^{-8}$-$9.29 \cdot 10^{-8}$. Neglecting mutations over frequency 0.46 does not alter the results. For illustration, Fig. 8 shows the VAF of a synthetic sample, generated using the tree fitting best the empirical data.

The estimated mutation rate is rather high, compared to estimations of the order of 10^{-9}-10^{-10} per site per cell division for healthy human somatic tissues [18]. In comparison with mutation rates of tumors, it is not an outstanding value [1]. Meanwhile, the turnover rate is also high, being very close to the birth rate. Possible causes include the effect of the immune system, the deleterious nature of driver mutations or competition for resources among tumor cells. In conclusion, for this tumor sample, the high number of mutations is due to a combination of an elevated mutation rate and a high turnover rate.

The results allow estimating the number of cell division rounds from the founding cell to the biopsied tumor. The average height of simulated trees with the estimated parameters is 2009 cell divisions. It should be noted that a naive

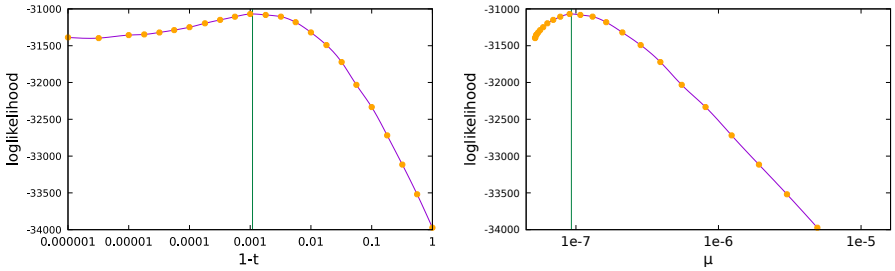

Figure 7: Loglikelihood-turnover rate and loglikelihood-mutation rate curves of the HCC data. Interpolation between data points is by cubic splines.

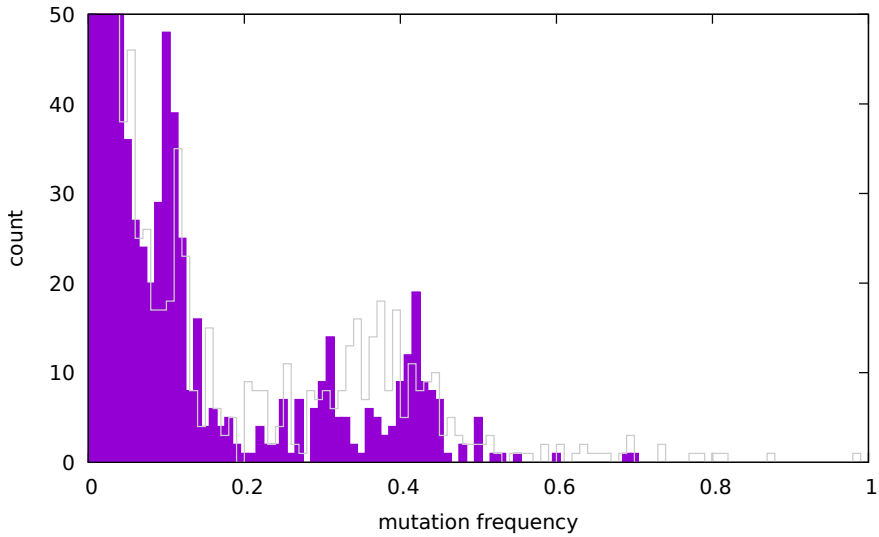

Figure 8: VAF of a synthetic sample, generated using the tree which fits the empirical data the best. The grey outline shows the empirical VAF.

estimation of tree height using $\log_2(2.7 \cdot 10^9)$ successive branches of average length $1/(1.1 \cdot 10^{-3})$ is wrong, due to the very different shapes of surviving trees compared to all trees, most of which go extict before reaching $1/(1.1 \cdot 10^{-3})$ size.

It is also possible to estimate the lifetime of the HCC sample and the cell division rate of the HCC tumor. The diameter of the tumor is 35 mm, while the length of a HCC cell is 25 μm [14]. This gives a total number of $2.7 \cdot 10^9$ cells in the whole tumor. The median HCC tumor volume doubling time is 86 days [19]. Based on these figures, the lifetime of the analysed sample is around 7 years, and the cell division rate is estimated to be around 1/32 1/hour.

5 Discussion

In summary, we described a method to simultaneously estimate the mutation rate and the turnover rate, making it possible to answer the question which of them is responsible for the elevated number of mutations in tumors. In particular, the mutated sites' read counts, which are closely related to the shape of the site frequency spectrum, contain useful information about the turnover rate (death rate, relative to the birth rate), even in the presence of a moderate sequencing error rate. The sequencing error rate can also be estimated. It is also quite straightforward to elaborate the model by including nucleotide-dependent transition probabilites, or trinucleotide context-based effects.

The accuracy of the estimation is influenced by 4 factors. First, the sharpness of the peak of the loglikelihood function, which tends to be narrow, for the expected amount of input data. Second, the finite amount of trees used in the fit, causing only a slight dispersion for 10^4 trees. Third, the shape of the true lineage tree, which, for death rates extremely close to the birth rate, can distort the estimation by one order of magnitude. Finally, the assumptions behind the model (birth-death process with constant rates, neutral mutations from a Poisson process) also contribute to the uncertainty of the estimations. According to the results, the estimation method works sufficiently well to discern cases of small difference between the birth and death rates ($\alpha - \beta \ll \alpha$) and cases of the death rate being much lower than the birth rate. Without such capability, the answer to the question "Is it the mutation rate or the death rate?" would always be "mutation rate".

Although the method is presented in the context of human tumors, it can handle healthy tissues, and samples from other species, too. In theory, any population, descending from one ancestor and possessing genetic material, can be analyzed, however, lengthier genomes giving rise to more mutations are easier, due to the increase in the input data for the estimation.

On the more practical side, we also discovered that averaging the loglikelihoods over trees, instead of the likelihoods, gives a significant improvement in the robustness of the results.

Concerning sequencing errors, the noise level in the standard Illumina technology makes applying the method to typical samples impractical. One solution is to use a sequencing technology with much lower error rates, e.g., [21, 22], or even below the 10^{-6} error rate of the PCR process, [13, 15]. It should be noted, however, that these technologies have been applied to short DNA segments only, resulting in a reduced number of mutations as input data. Another possibility is to apply noise filtering to standard sequencing data, e.g., deepSNV [23], and modify the error model of the fitting process accordingly.

Furthermore, when there is no estimation of the order of magnitude of the sequencing error rate, variance in the accuracy of the method can be quite large in individual cases, despite the much better behavior of the averaged results.

Despite the shortcomings, it is clear that the signal does exist in the site frequency spectrum, the mean of the estimated turnover rates monotonically changes with the testing datasets true turnover rates, and are clearly not independent of them.

Besides successes on synthetic data, we were also able to analyze an empirical sample of a hepatocellular carcinoma. We simultaneously estimated the mutation rate and the turnover rate. Both quantities were estimated to be much

higher than for healthy tissues, mutation rate being $9.3 \cdot 10^{-8}$ per site per cell division, and turnover rate $t = 0.9989$. In other words, the high number of mutations in this tumor is caused by a combination of both high mutation and turnover rates. Using the turnover rate, we also estimated the number of cell division rounds in the tumor's lifetime, and its cell division rate. The results suggest that tumor cells are constantly dividing, but the growth of the tumor is limited by other factors changing on a much longer timescale, e.g., securing sufficient blood supply, which cause most new tumor cells to die and slow down tumor growth. With such a high turnover rate, the ability of limitless replication is essential for tumor growth.

It is interesting to note that high turnover rates are able to reproduce subclonal peaks in the VAF, using a purely neutral birth-death process. In other words, subclonal peaks are not necessarily the consequence of selection, neutral processes can also produce them, indicating strong cell death. On this basis, it is possible to give a definition of subclones as branches in the lineage tree, close to the root and long enough to appear as peaks on the VAF spectrum. Using this definition, there is no need to explain different parts of the VAF spectrum with different models[24].

In this work, the turnover rate was held constant during the evolutionary process. There are signs that it is more realistic to assume a turnover rate which changes during tumor growth [2]. In our case, the estimated strong cell death suggests that the tumor reached a slowly growing phase, in line with a Gompertzian model of tumor growth [25, 26], which is corroborated by the large sizes of observed tumors (diameter $\geq$ 1cm) used in the doubling time estimation [19]. It is possible that in the earlier stages of tumor development, cell death was less frequent and doubling time was shorter. It might be the case that the rate of cell division is constant during tumor growth, and doubling time is set by the turnover rate, which is, in turn, limited by external factors. It is an interesting direction for future work to extend the model by allowing the turnover rate to be time-dependent.

The combined effect of the estimated mutation and turnover rates is a very high effective mutation rate between cell divisions where both daughter branches survive, μ_{eff} being in the order of 10^{-4}-10^{-5} per site per surviving cell division. While this value looks suprisingly large, it is logical that the combination of a slowly growing tumor and fast dividing tumor cells leads to a very large number of mutations.

Currently, the method uses a simple birth-death model for tumor growth. In the future, a more realistic growth model, including e.g., spatial effects [27, 28], would enhance the applicability of the method. Another possibility for improvement is to model spatial sampling of tissues, in which the measured mutation frequencies intertwine the correlated ancestry of sampled cells with the prevalence of the mutations.

Author Contributions

(According to https://journals.plos.org/ploscompbiol/s/authorship#loc-author-contributions.)

GJSz conceptualized the research project. GT, ID and GJSz performed the formal analysis. Funding was acquired by ID and GJSz. GT carried out the

investigation. GT, ID and GJSz contributed to the methodology. GT and DS developed the necessary software. GT provided the visualization. GT wrote the draft. DS, ID and GJSz reviewed and commented on the manuscript.

Acknowledgements

GT and GJSz received funding from the European Research Council under the European Unions Horizon 2020 research and innovation programme under grant agreement no. 714774. GJSz was also supported by the grant GINOP-2.3.2.–15–2016–00057.

References

[1] Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Gen. 2016;48:238–244.

[2] Bozic I, Gerold JM, Nowak MA. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution PLoS Comput Biol. 2016;12(2):e1004731.

[3] Tomlinson I, Sasieni P and Bodmer W. How Many Mutations in a Cancer? Am J Pathol. 2002;160:755-758.

[4] Araten DJ, Golde DW, Zhang RH, Thaler HT, Gargiulo L, Notaro R et al. A Quantitative Measurement of the Human Somatic Mutation Rate. Cancer Research 2005;65:8111.

[5] Loeb LA, Bielas JH and Beckman RA. Cancers Exhibit a Mutator Phenotype: Clinical Implications. Cancer Research 2008;68:3551.

[6] Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nat Genet. 2018;50:895-901.

[7] Werner B, Case J, Williams MJ, Chkhaidze K, Temko D, Fernández-Mateos J, et al. Measuring single cell divisions in human tissues from multi-region sequencing data. Nat Comm. 2020;11:1035.

[8] Gernhard T. The conditioned reconstructed process. J Theor Biol. 2008;253:769.

[9] Maruvka YE, Kessler DA, Shnerb NM. The Birth-Death-Mutation Process: A New Paradigm for Fat Tailed Distributions. PLoS ONE 2011;6(11):e26480.

[10] Kessler DA, Levine H: Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. J Stat Phys. 2015;158:783–805.

[11] Schrempf D. The ELynx Suite; 2019 [cited 2020 Sept 01] Repository: GitHub [Internet]. Available from: https://github.com/dschrempf/elynx

[12] Höhna S, May MR, Moore BR. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics 2016;32(5):789-791.

[13] Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9:2586.

[14] Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, et al. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. PNAS 2015;112:E6496-E6505.

[15] Wang K, Lai S, Yang X, Zhu T, Lu X, Wu C, et al. Ultrasensitive and high-efficiency screen of de novo low-frequency mutations by o2n-seq. Nat Comm. 2017;8:15335.

[16] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-76.

[17] Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics 2017;18:286.

[18] Lynch M. Evolution of the mutation rate. Trends Genet. 2010;28:345.

[19] An C, Chou YA, Choi D, Paik YH, Ahn SH, Kim M-J, et al. Growth rate of early-stage hepatocellular carcinoma in patients with chronic liver disease. Clin Mol Hepatol. 2015;21:279.

[20] Stadler T. On incomplete sampling under birth-death models and connectsion to the sampling-based coalescent. J Theor Biol. 2009;261:58-66.

[21] Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. PNAS 2013;110:19872-19877.

[22] Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. PNAS 2011;108:9530-9535.

[23] Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Comm. 2012;3:811.

[24] Caravagna G, Heide T, Williams MJ , Zapata L, Nichol D, Chkhaidze K, et al. Subclonal reconstruction of tumors by using machine learning and population genetics. Nat Gen. 2020;52:898–907.

[25] Laird AK. Dynamics of Tumour Growth: Comparison of Growth Rates and Extrapolation of Growth Curve to One Cell. Br J Cancer 1965;19:278–291.

[26] Lo CF. A modified stochastic Gompertz model for tumour cell growth. Comp Math Methods Medicine 2010;11:3-11.

[27] Antal T, Krapivsky PL, Nowak MA. Spatial evolution of tumors with successive driver mutations. Phys Rev E 2015;92:022705.

[28] Noble R, Burri D, Kather JN, Beerenwinkel N. Spatial structure governs the mode of tumour evolution. BioRxiv [Preprint]. 2019 bioRxiv 586735 [posted 2019 Mar 23; revised 2019 Apr 13, cited 2020 Sept 24]: [19 p.]. Available from: https://doi.org/10.1101/586735

[29] Nelso KP. Assessing Probabilistic Inference by Comparing the Generalized Mean of the Model and Source Probabilities. Entropy 2017;19:286.

[30] Nelso KP. Inference assessment on a probability scale. 51st Annual Conference on Information Sciences and Systems (CISS) 2017; 1-5.10.1109/CISS.2017.7926106.