Distinguishing excess mutations and increased cell death based on variant allele frequencies

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

[1]MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Eötvös Loránd Research Network (ELKH), 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, Eötvös Loránd Research Network (ELKH), Pázmány P. stny. 1A, H-1117 Budapest, Hungary

July 23, 2021

Abstract

Tumors often harbor orders of magnitude more mutations than healthy tissues. The increased number of mutations may be due to an elevated mutation rate or frequent cell death and correspondingly rapid cell turnover, or a combination of the two. It is difficult to disentangle these two mechanisms based on widely available bulk sequencing data, where mutations from individual cells are intermixed and, thus, 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. Our method works by simulating tumor growth and finding the parameters with which the observed data can be reproduced with maximum likelihood. Applying this method to a real tumor sample, we find that both the mutation rate and the frequency 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 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, but only becomes apparent in late stages and, as a result,

is most often not directly observed during large part of its evolution. Genomic sequencing data offers a window into the evolutionary processes underlying tumor development and progression. Analyzing widely available bulk sequencing data, where mutations from individual cells are intermixed, however, is challenging. Bulk sequencing can only essay mutation frequencies for a population of cells from each tumor sample and does not resolve the genotypes of individual cells, making even basic evolutionary parameters, like mutation rate and death rate, difficult to recover.

Consider the neutral model of tumor growth, where new mutations can appear with each cell division, while cells can also die for reasons such as lack of nutrients or immune reactions [1, 2]. This simple process can be described by two paramaters, the ratio of cell death and birth rates and the mutation rate per cell division. Despite the simplicity of such a model and a plethora of bulk tumor sequencing data, these parameters remain largely unknown, with estimates spanning several orders of magnitudes [2].

While tumor cells can contain a large number of mutations, it is not clear whether this is due to an elevated mutation rate or frequent cell death and birth. There are arguments for both cases [3, 4, 5, 6, 7], but distinguishing between these two alternatives is difficult because there are no known methods for direct measurements. As a result, estimating the mutation rate per cell division in human tumors must rely on assumptions about the duration of the cell cycle, the growth rate of the tumor and the total mutational burden of the tumor [8, 9, 10].

In previous works, Williams et al. [2] and Bozic et al. [3] showed that the combined effect of the mutation rate and the death rate can be estimated from the frequencies of neutral mutations that is readily available from bulk sequence data. Distinguishing excess mutations and increased cell death, however still required external information or assumptions. In follow up work, Williams et al. [11] showed that separating these two quantities can be achieved by the bulk sequencing of multiple spatially disjunct samples from the same tumor, thus, resolving a coarse grained cell linage tree. This approach, however, is inherently limited by the number of samples in its ability to resolve the cell lineage tree and, as a result, in its ability to distinguish excess mutations from increased cell death.

Here, we describe a method for simultaneously estimating the per-cell mutation rate and the death-to-birth ratio of cells (characteristic of cell turnover) of a tumor from bulk sequencing data. The method estimates the parameters of a birth-death process from mutant and wild-type read counts from bulk sequncing data. Despite requiring only a single tumor-normal sample pair it can differentiate between death rates that are extremely close to the birth rate, resulting in accurate estimation of mutation rates over orders of magnitudes. The rest of the paper is structured as follows. After introducing our model and the fitting procedure we assess the accuracy of our method on simulated data and describe results on empirical data.

2 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 we consider the tree spanned by surviving cells. The leaves of this tree correspond to extant cells and the internal nodes to cell divisions. To model the descendance of the extant cells, we employ the conditioned birth-death process with birth rate (which is identical to the cell division rate and is also referred to as the cell turnover rate) b, death rate d, and number of sampled cells n [12]. We measure branch lengths in number cell divisons, i.e., birth events. Consequently, the birth rate b can be considered as a scaling constant that sets the unit of time. Changing the unit of time - an arbitrary decision of ours - should not change the shape of the tree. Therefore, rescaling the rates should have no effect. Thus, the relevant parameter that determines the structure of the cell lineage tree is the death-to-birth ratio $\delta = d/b$. We only consider exponentially growing cell populations, i.e., $\delta < 1$. Mutations occur at a rate μ per site per cell division. They are considered neutral and we neglect the probability that a site is hit by mutation more than once.

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 data available from bulk sequencing are the mutant and wild type read counts at each site. To illustrate how such data can be used to separate the effects of the mutation rate and cell death intensity, consider the variant allele frequency (VAF) spectrum, which can be estimated from read count data. As shown in (Fig. 1a and b) the site frequency spectrum reflects the branch length distribution of the tumor's cell lineage tree, the leaves of which are the sequenced cells, and its root is the most recent common ancestor of these cells. Changing the death-to-birth ratio modifies the shape of the cell linage tree by changing the relative lengths of branches closer to and further away from the root and, as a result, modifies the site frequency spectrum. Changing the mutation rate, on the other hand, simply changes the number of mutations uniformly along the tree, thus, leaves the overall shape of the site frequency spectrum intact (Fig. 1c and d).

It should be noted, however, that sequencing data contains more information than the VAF spectrum. For each site the mutant and total read counts, and not only their ratio is known (e.g., 10 mutant reads out of 30 total reads have different statistical significance from 1 out of 3, even though they both correspond to a mutation frequency of 1/3).

To compare different combinations of mutation rates and death-to-birth ratios describing the observed empirical data we employ a likelihood based approach. First, we derive the likelihood $\mathcal{L}(D|\mu, \delta)$ of the observed data D as a function of the per site mutation rate μ and the death-to-birth ratio δ. As described below we maximize a score based on the likelihood function using random samples of independent and identically distributed cell lineage trees, in order to estimate the parameters 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 $\mathcal{T}$. We assume that sites collect mutations independently of each other. Consequently, $\mathcal{L}(D|\mu, \mathcal{T})$ takes the form

$$\mathcal{L}(D|\mu, \mathcal{T}) = \prod_i p(m_i|\mu, \mathcal{T}, r_i) \tag{1}$$

where the product runs over all sites, m_i is the number of reads exhibiting a

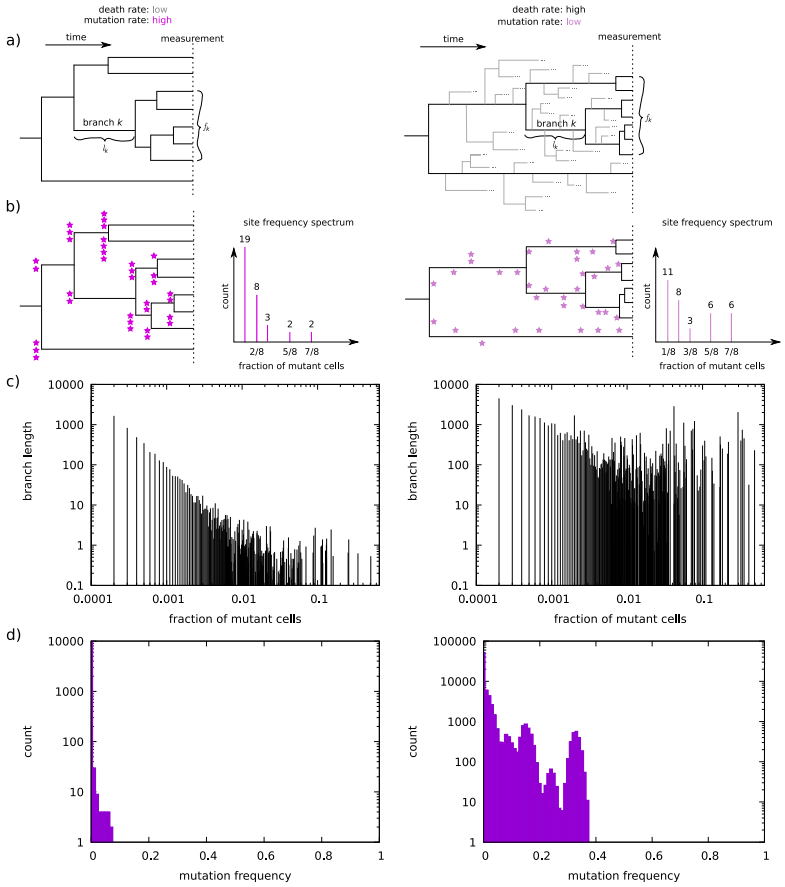

Figure 1: **Distinguishing mutation rate and cell death intensity based on variant allel frequencies.** Two possible scenarios for the generation of mutations along cell lineage trees. **a)**: different death-to-birth ratios lead to different lineage tree shapes. Bifurcations are cell divisions, leaves 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 death-to-birth ratio and low mutation rate can lead to the same number of observed mutations as low death-to-birth ratio and high mutation rate, however, the mutation spectrum of the two trees are different. **c)**: For simulated trees of 10^4 leaves, the differences in the branch length distribution are clearly visible. **d)**: VAF spectra for death-to-birth ratios of 0 (left panel) and 0.999999 (right panel). Fractions of mutant cells are binned (note the logarithmic scale). Ploidy is set to two, contamination is zero. Simulated sequencing depth is 1000. The VAF spectra are based on the trees used to generate subplot **c)**.

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 two alternatives: First, if $m_i = 0$, then either a mutation occurred with probability $\mu \cdot l_k$ on some branch k with length l_k, but no mutant

read was observed out of the r_i reads, or with probability $1-\mu\cdot\sum_k l_k = 1-\mu\cdot L$ no mutation occurred on any of the branches. Second, if $m_i > 0$, then a mutation occurred on some branch k with length l_k with probability $\mu\cdot l_k$ and m_i mutant reads were observed out of r_i. Thus,

$$p(m_i|\mu,\mathcal{T},r_i) \;=\; \begin{cases} \sum_k \mu\cdot l_k\cdot \mathrm{Binom}(0,r_i,f_k) + (1-\mu\cdot L)\,, & m_i = 0 \\ \sum_k \mu\cdot l_k\cdot \mathrm{Binom}(m_i,r_i,f_k), & m_i > 0 \end{cases} \tag{2}$$

where $\mathrm{Binom}(m,r,f)$ is the binomial distribution of m successes out of r independent Bernoulli trials with success probability f, and f_k denotes the fraction of sequenced leaf cells that descend from branch k. Multiple mutations at the same site are neglected, which is an appropriate approximation if $\mu\cdot L \ll 1$. In all of our applications we verified that this condition is satisfied.

To take into consideration sequencing errors, we must consider that they lead to an excess of spurious mutant reads that are in fact the result of sequencing error. We do this by introducing the probability ε of a sequencing error per site per read. In the presence of sequencing error (i.e., $\varepsilon > 0$) a read can be i) an apparent mutant read caused by an actual mutation (a true mutant read) or a sequencing error (a false mutant read), alternatively it can be ii) an apparent wild type read due to no mutation (a true wild type read) or because a mutation was reverted by a sequencing error (false wild type read):

$$\begin{aligned} p(m_i|\mu,\mathcal{T},r_i,\varepsilon) \;\approx\; & \sum_k \mu\cdot l_k\cdot \mathrm{Binom}\left[m_i,r_i,f_k\cdot(1-\varepsilon)+(1-f_k)\cdot\varepsilon\right] \\ & + (1-\mu\cdot L)\cdot \mathrm{Binom}(m_i,r_i,\varepsilon). \end{aligned} \tag{3}$$

Note that this equation incorporates both the $m_i = 0$ and $m_i > 0$ cases.

So far, we have assumed that each site can have 2 states, wild type or mutant, corresponding to a DNA consisting of only two types of nucleotides, rather than four. It is, however, straightforward to introduce multiple mutant types. The likelihood function for 3 possible mutant types relevant for DNA sequences that we use for inference where indicated, can be found in the Appendix.

Parameter inference

A conceptually straightforward approach for treating the unknown cell linage tree as a nuisance parameter is to average over all trees $\mathcal{T}$ [13] according to their probability:

$$\mathcal{L}(D|\mu,\delta) = \sum_{\mathcal{T}} \mathcal{L}(D|\mu,\mathcal{T})\cdot p_{\mathrm{BD}}(\mathcal{T}|\delta), \tag{4}$$

where $p_{\mathrm{BD}}(\mathcal{T}|\delta)$ is the probability of the cell lineage tree $\mathcal{T}$ conditioned on a birth-death process with death-to-birth ratio δ. Due to the very large number of possible trees the above average is intractable and because of the inherent lack of resolution resulting from bulk sequencing data a Markov Chain Monte Carlo sampler is impractical. To overcome these issues, here we resort to sampling a finite number of trees from the conditioned birth-death process with fixed δ. However, as we can sample only a small fraction of trees, the estimated likelihood is typically dominated by a single tree, hence, the estimation becomes sensitive

to the particular realization of the sample set. Using synthetic data (see below), however, we demonstrate that maximizing the average of the log-likelihood

$$\ln \bar{\mathcal{L}}(D|\mu,\delta) = \frac{1}{|\mathcal{T}(\delta)|} \sum_{\mathcal{T}(\delta)} \ln \mathcal{L}(D|\mu,\mathcal{T}), \tag{5}$$

where $|\mathcal{T}(\delta)|$ is the number of sampled trees, allows accurate and robust inference that is less sensitive to sampling noise.

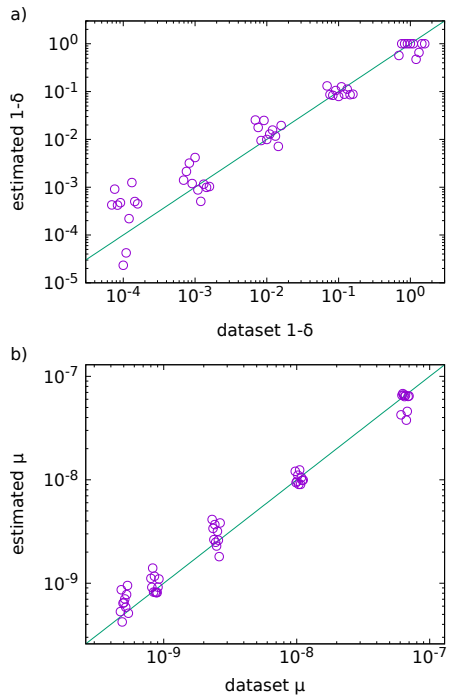

Figure 2: **True vs. estimated death-to-birth ratios and mutation rates** On subplot a), one minus death-to-birth ratio $(1-\delta)$ is shown for better visualization. On both a) and b), 10 synthetic datasets were generated for each true value. 10^4 trees with 10^4 leaves were used for fitting (see the Appendix for details on calculating the likelihood). Points are slightly dispersed horizontally for clarity. Datasets are the same for a) and b). 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.

The mutation rate (defined as the number of mutations divided by the number of sites and the total branch length of the tree) is estimated directly from the data, in order to speed up the inference. Since some mutations are expected to be sequencing errors, we only count mutations of significant read count. The threshold number of mutant read counts is set in a sequencing depth dependent manner such that the expected number of sites with mutant reads that result from sequencing error is less than one for the entire dataset by choosing:

$$m_{\text{th}}(r,\varepsilon) = \min\{m : \text{Binom}(m,r,\varepsilon) < \frac{1}{n_{\text{sites}}}, m > \varepsilon \cdot r\}, \tag{6}$$

where r is the number of reads (i.e., the local sequencing depth) and n_{sites} is the total number of sites in the data. The expected total branch length of $\mathcal{T}$ corresponding to the significant mutant read counts can be estimated as

$$L_{\text{sig}}(\mathcal{T}, \varepsilon) = \sum_{r,k} \frac{n_r}{n_{\text{sites}}} \cdot l_k \cdot \sum_{m' \geq m_{\text{th}}(r,\varepsilon)} \text{Binom}(m', r, f_k), \tag{7}$$

where n_r is the number of sites with r reads and n_{sites} is the genome length (is number of all the sites). Using the above formula, the mutation rate can be estimated as:

$$\mu_{\text{est}}(\delta, \varepsilon) = \mathbb{E}\left[\frac{M_{\text{sig}}(\varepsilon)}{n_{\text{sites}} \cdot L_{\text{sig}}(\mathcal{T}, \varepsilon)}\right]_{\mathcal{T}(\delta)}, \tag{8}$$

where the $M_{\text{sig}}(\varepsilon) = \sum_{i, m_i \geq m_{\text{th}}(r_i, \varepsilon)} 1$ is the number of sites that have at least a threshold number of mutant reads in the data.

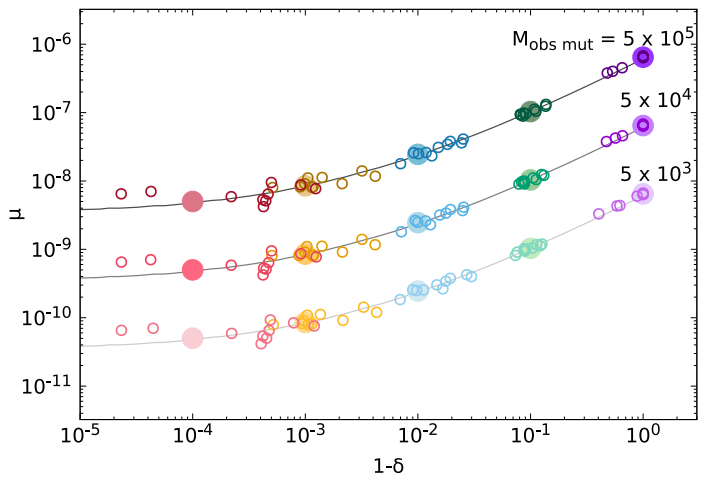

Figure 3: **Joint estimates of mutation rate and death-to-birth ratio pairs.** 10 synthetic dataset replicates were generated for each true parameter pair, with replicates denoted by the same color. True parameter values are indicated by large filled circles. Solid lines show the numerical approximation of parameter pairs for different values of $M_{\text{obs mut}}$. Data around the middle line are the same as on Fig. 2.

3 Results on synthetic data

We validated the parameter estimation method on simulated data generated according to the model described above. The simulation procedure is described in the Appendix.

No sequencing errors

Figs. 2 shows how the estimated death-to-birth ratios and mutation rates compare to the true death-to-birth ratios and mutation rates. The method can rea-

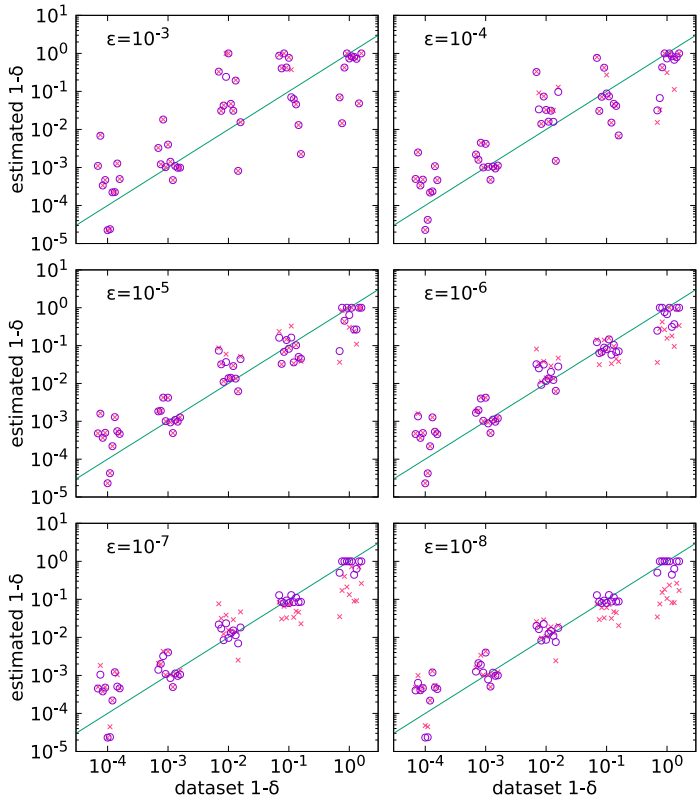

Figure 4: **The effect of varying the error rate.** Sequencing error rates are shown in the subplots. 10^4 trees with 10^4 leaves were used for fitting. Horizontal coordinates are slightly dispersed for clarity. Open circles are results corresponding to error rates fixed to their true values, crosses correspond to error rates estimated by the parameter fit. Each open circle-cross pair corresponding to the same dataset is vertically aligned.

sonably differentiate between datasets with different death-to-birth ratios and mutation rates, and estimate their values.

Fig. 3 shows the joint estimation of mutation rate and death-to-birth ratio pairs. The simulated data points are grouped into isolines with a constant number of observed mutations

$$M_{\text{obs mut}}(\mu, \delta) = \mathbb{E}\left[\mu \cdot \sum_k l_k \cdot (1 - \text{Binom}(0, \overline{r}, f_k))\right]_{\mathcal{T}(\delta)}, \tag{9}$$

where $\overline{r}$ denotes the average sequencing depth (average read count). The expected value is approximated by averaging over trees generated with death-to-birth ratio δ.

Results on simulated data indicate that the accuracy of parameter estimates for different datasets are not uniform. As Fig. S4 shows resolving large death-to-birth ratios requires trees with more leaves, and accuracy is increased when analyzing trees with a sufficiently large number of leaves. Aside of the effect of

the size of the trees in case of high death-to-birth ratios, differences between estimates for different datasets can also result from: (i) sampling noise from the trees used in Eq. (5), (ii) stochasticity of the simulated data and (iii) the simulated cell lineage tree used to generate the read counts. To differentiate between these possible factors, we chose a dataset with an estimated death-to-birth ratio that markedly deviated from the true value (dataset no. 7 for $1-\delta = 1.0$ in Fig. 2 with an estimated value of $1-\delta = 0.47$). We calculated the estimated death-to-birth ratio values using 10 independent sets of 1000 trees. Estimates were obtained in the range of $1-\delta \in [0.39, 0.54]$. Therefore, the deviation of the estimate from 1.0 cannot be attributed to the sample of fitting trees. Then, we simulated 10 additional datasets using the same tree as for the original dataset. The estimated death-to-birth ratio values were: $1-\delta \in [0.44, 0.49]$, even more closely matching the original estimate. Consequently, the effect does not depend on the simulated mutations but on the tree along which they were generated. This suggests that the deviation of the estimates from the true death-to-birth ratios primarily reflects the fluctuation of the shapes of the trees used for sample generation. It seems that estimating the distribution of bifurcation times of the generating tree is not an easy task.

Effects of sequencing error

To estimate the effect of sequencing errors we re-estimated death-to-birth ratios while varying the magnitude of sequencing error, but using the same initial read counts as in Fig. 2). In this case, the error rate of the data was also estimated by the fitting procedure, along with the mutation rate and the death-to-birth ratios. The range of the error rates is from 10^{-3}, which is cited as the error rate of the Illumina sequencing technology [14], to 10^{-8}, which is what advanced technology can achieve [15]. The influence of sequencing errors on the estimation of the death-to-birth ratios is shown in Fig. 4. For an error rate of 10^{-3} (the case of the Illumina technology), the estimated death-to-birth ratios can have significant deviations from the true values.

4 Results on empirical data

To estimate *in vivo* death-to-birth ratios and mutation rates, a tumor sample is required. Due to the sensitivity of our method to high sequencing error rates, we need a sample which is sequenced using a very low error rate technology. We found a sample of a human hepatocellular carcinoma (HCC) [16], which was sequenced using the o2n sequencing technology [15], providing error rates between 10^{-5} and 10^{-8}, which is significantly lower than the 10^{-3} rate of the standard Illumina process. Besides the low error rate, the dataset also has very high sequencing depth coverage, the average sequencing depth is 904, over 923383 sites. High sequencing depth allows the identification of more mutations and provides better resolved VAF spectra.

After preprocessing the data (see details in the Appendix), 2284 mutations were identified. The variant allele frequency spectrum is shown in Fig. 5. Trees of 10^4 leaves were used for estimating the parameters, as DNA was extracted and quantified from 10^4 tumor cells that were precisely collected by laser capture microdissection (LCM) [16]. Using the procedure described above we obtained: a

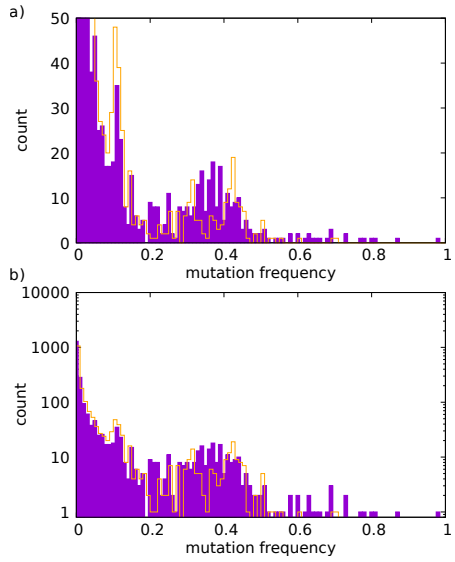

Figure 5: **Estimating the death-to-birth ratios and mutation rates on empirical data. a)** The solid purple bars show the VAF spectrum of a human hepatocellular carcinoma sample [16], obtained by o2n sequencing [15]. The orange line shows the VAF spectrum of a synthetic sample, generated using the tree having the highest likelihood for the empirical data. The tree has 10^4 leaves corresponding to the number of sequenced tumor cells. **b)** counts are shown on a logarithmic scale.

sequencing error of $\varepsilon = 10^{-7}$, consistent with the reported error rate of between 10^{-5} and 10^{-8} [15] for o2n, a death-to-birth ratio of $\delta = 1 - 1.1 \times 10^{-3}$, and a mutation rate of $\mu = 9.3 \times 10^{-8}$ per site per cell division. Fig. 5 also shows the VAF of a synthetic sample, generated using the sample tree that fits best the empirical data. Fig. S2 in the SI shows that the mutation rate and the death-to-birth ratio do not depend strongly on the value of the error rate.

Estimations for the mutation rate of healty somatic cells can come from fitting mathematical models to the development of certain tumors, or cell line experiments, both methods observing mutations on a single gene [17]. A third method is to count the number of mutations in sequenced cells and estimate the number of corresponding cell divisions [18]. Finally, a mathematical model fitted to multiple samples from the same donor was applied to estimate mutation rates during early development in haematopoietic stem cells and neurons [11]. For all of these studies, estimated mutation rate values are in the range 10^{-9} to 10^{-10} per site per cell division [17, 18, 11], lower than our estimation of the hepatocellular carcinoma data.

Estimations of the mutation rate in tumors have high variance. In colorectal cancers, $\mu = 5 \times 10^{-10}$ per site per cell division [8] was estimated by observing the number of mutations and estimating the corresponding number of cell divisions. In kidney cancer, $\mu = 2 \times 10^{-9}$ per site per cell division was estimated [11], using multiple samples from the same tumor and fitting a mathematical model. The same study estimated $\mu = 5 \times 10^{-8}$ per site per cell division for lung cancer.

An effective mutation rate was estimated using a different mathematical model, in which the mutation and death rates are not separable [2]. Assuming no cell death ($\delta = 0$), for brain and prostate cancers $\mu = 9 \times 10^{-8}$ per site per cell division was estimated, along with $\mu = 8 \times 10^{-7}$ per site per cell division for lung and bladder cancers. Substituting our estimation of $\delta = 0.999$, these change to $\mu = 9 \times 10^{-11}$ and 8×10^{-10} per site per cell division, respectively. Our estimation of $\mu = 9.3 \times 10^{-8}$ per site per cell division for HCC is at the higher end of the estimations for the mutation rates in tumors.

For the death-to-birth ratio, earlier esimations range within $\delta \in [0.08, 0.97]$ [11] for colon cancer, using the same methodology as above. In another study [3], $\delta = 0.72$ was estimated for fast-growing colorectal cancer metastases and $\delta = 0.999$ for premalignant colorectal tumors. Upper limits are provided [2] within $\delta \in [0.99, 0.999]$, assuming $\mu = 10^{-9}$ per site per cell division. Using our estimated mutation rate $\mu = 9.3 \times 10^{-8}$ per site per cell division, this changes to $\delta \in [0, 0.88]$. Our estimated death-to-birth ratio is similar to the previously estimated $\delta = 0.999$ of premalignant colorectal tumors. Possible causes for such an elevated death rate include the effect of the immune system, the deleterious nature of 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 death-to-birth ratio.

The above results allow us to estimate the number of cell division rounds from the founder cell to the biopsied tumor. The average height of the simulated trees with the estimated parameters is 2023 cell divisions. This value might seem counterintuitively low at first, according to the following estimation. A naive estimation of the average tree height would assume that to reach the size of $n_{\text{tumor cells}}$ leaves, the tree should have $\log_2 n_{\text{tumor cells}}$ branches between the root and a leaf. The average length of such a branch should be $1/(1-\delta) = 1/(1.1 \times 10^{-3})$. As $n_{\text{tumor cells}} = 2.7 \times 10^9$ (see the next paragraph), the naive estimation of tree height is $\log_2(2.7 \times 10^9) \cdot (1/1.1 \times 10^{-3}) = 2.8 \times 10^4$. This reasoning, however, does not take into account the fact that trees with a death-to-birth ratio close to 1 have very different shapes compared to cell lineage trees resulting from a pure birth process. Consequently, the much lower tree height of 2023 as opposed to 2.8×10^4 can be attributed to the difference between the average tree with birth-to-death ratio $1 - 1.1 \times 10^{-3}$ and the average tree of a pure birth process stretched out by a factor of $1/(1.1 \times 10^{-3})$.

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 [16]. This gives a total number of 2.7×10^9 cells in the entire tumor. The median HCC tumor volume doubling time is 86 days [19]. Based on these figures, the lifetime of the analyzed sample is around 7 years, and the cell division rate is estimated to be around $2023/[\log_2(2.7 \times 10^9) \cdot 86\text{days}] \approx 0.75\ \text{day}^{-1}$.

In the above calculation, the effect of sampling was neglected, which corresponds to cutting out exactly the descendants of one branch in the tumor lineage tree. This can be regarded as an approximation of taking a local sample from a solid tumor, and the estimated parameters refer to that branch of the lineage tree. We note that the corresponding division rate is realistic, suggesting that the approximation is adequate.

Random sampling from a perfectly mixed population can be considered as an opposite extreme. In this case, the transformation of the birth rate is known

[20]:

$$b = b'/\rho, \tag{10}$$

where ρ is the sampling ratio and b' is the birth rate corresponding to the $\rho = 1$ case. In our case, the size of the tumor is 2.7×10^9 cells, from which 10^4 cell were sampled. Then the sampling ratio is $\rho = 10^4/(2.7 \times 10^9) = 3.7 \times 10^{-6}$ and the division rate is $b = 2.03 \times 10^5$ day^{-1}, which is biologically not realistic. This estimation suggests that the sampling ratio may be close to 1, as the corresponding division rate is much more realistic.

5 Conclusion

In summary, we describe a method to simultaneously estimate the mutation rate and the death-to-birth ratio together with the sequencing error rate, making it possible to answer the question which of the two 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 death-to-birth ratio, even in the presence of a moderate sequencing error rate.

Our results on simulated data show that the estimation method can resolve the death-to-birth ratio even if the birth and death rates are close to each other (i.e., $1 - \delta \ll 1$) as long as the sequencing error is sufficiently small. Unfortunately, for the error rates of standard Illumina sequencing, the estimation has a significant variance, therefore applying the method to typical samples is impractical. One solution is to use a sequencing technology with much lower error rates, e.g., as in Refs. [21, 22], or even below the 10^{-6} error rate of the PCR process [14, 15]. 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.

As a proof of concept application we analyzed low error sequencing data from human hepatocellular carcinoma [15]. For both the death-to-birth ratio and the mutation rate we recovered estimates, a mutation rate being 9.3×10^{-8} per site per cell division, and death-to-birth ratio $\delta = 0.9989$, which are higher than expected for healthy tissues but fall within the range of previous estimates for tumors [2].

It is important to note that high death-to-birth ratios can produce marked subclonal peaks in the VAF spectrum (Fig. 5). This implies that subclonal peaks are not necessarily the consequence of selection. Neutral processes, in particular a high death-to-birth ratio, can also produce such signal.

In this work, the death-to-birth ratio was assumed to be constant during the evolutionary process. It is more realistic to assume a death-to-birth ratio which changes during tumor growth [3]. 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 [24, 25], 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 earlier stages of tumor development, cell death is less frequent and doubling time is shorter. It might be the case that the rate of cell division is constant during tumor growth, and doubling time is set by the death-to-birth ratio, which, in turn, is limited by external factors. It is an interesting

direction for future work to extend the model by allowing the death-to-birth ratio to be time-dependent.

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 [26, 27], 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.

Appendix

3 mutant types

For 3 possible mutant types, relevant for DNA sequences, 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 need to contain three mutant read counts rather than one for each site. 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 mutation at the same site is still neglected, because it is very rare, technically a second-order process in the mutation probability of a site. We also neglect the probability of more than one error hitting the same site. The likelihood function at a single site is then

$$\begin{aligned}
&p(m_i^{(1)}, m_i^{(2)}, m_i^{(3)} | \mu, \mathcal{T}, r_i, \varepsilon) \approx \\
&\quad \approx \sum_k \mu \cdot l_k \cdot \frac{1}{3} \sum_{j=1}^{3} \mathrm{Mult}\Bigg(\Big(m_i^{(j)}, m_i^{(j+1)}, m_i^{(j+2)}, r_i - \sum_{j'} m_i^{(j')} \Big); \\
&\qquad r_i;\ \Big(p_{\mathrm{M}}^{(j)}(f_k, \varepsilon), p_{\mathrm{M}}^{(j+1)}(f_k, \varepsilon), p_{\mathrm{M}}^{(j+2)}(f_k, \varepsilon), p_{\mathrm{W}}(f_k, \varepsilon) \Big) \Bigg) + \\
&\qquad + (1 - \mu \cdot L) \cdot \mathrm{Mult}\Bigg(\Big(m_i^{(1)}, m_i^{(2)}, m_i^{(3)}, r_i - \sum_{j} m_i^{(j)} \Big); \\
&\qquad\qquad r_i;\ \Big(p_{\mathrm{M}}^{(j)}(0, \varepsilon), p_{\mathrm{M}}^{(j+1)}(0, \varepsilon), p_{\mathrm{M}}^{(j+2)}(0, \varepsilon), p_{\mathrm{W}}(0, \varepsilon) \Big) \Bigg)
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
p_{\mathrm{M}}^{(j)}(f_k, \varepsilon) &= f_k \cdot (1 - \varepsilon) + (1 - f_k) \cdot \varepsilon/3 \\
p_{\mathrm{M}}^{(j+1)}(f_k, \varepsilon) &= f_k \cdot \varepsilon/3 + (1 - f_k) \cdot \varepsilon/3 = \varepsilon/3 \\
p_{\mathrm{M}}^{(j+2)}(f_k, \varepsilon) &= f_k \cdot \varepsilon/3 + (1 - f_k) \cdot \varepsilon/3 = \varepsilon/3 \\
p_{\mathrm{W}}(f_k, \varepsilon) &= f_k \cdot \varepsilon/3 + (1 - f_k) \cdot (1 - \varepsilon)
\end{aligned} \tag{12}$$

where (j), $(j+1)$, and $(j+2)$ denote the three possible mutant types with cyclic notation $(j) = (j+3)$, and Mult(vector of numbers of outcomes; number of trials; vector of probabilities of outcomes) is the multinomial distribution. The factor of 1/3 is due to the fact the if a mutation occurs at a site than the probability of each of the three possibilities is 1/3. In the multinomial distribution the first outcome corresponds to the true mutant type, the second and third to the false mutant types, and the forth to the wild type.

Generating trees

Cell lineage trees were generated by the ELynx software suite [28], using conditioned birth-death processes. Birth rates, death rates, and tree sizes are freely adjustable with this software.

Generating synthetic samples

Simulated data consists of mutant and wild type read counts based on mutations generated along a cell lineage tree $\mathcal{T}$ obtained using the ELynx suite.

Mutations are generated as follows. For each site, each branch k is checked for contributing a mutation with probability $1 - e^{-\mu l_k} \approx \mu l_k$. The first branch providing a hit is selected.

The total number of reads of the current site is drawn from a prescribed distribution.

The number of mutant reads is drawn form a hypergeometric distribution, corresponding to the branch frequency f_k. Sequencing errors, if required, are introduced using a multinomial distribution, with probabilities $\varepsilon/3, \varepsilon/3, \varepsilon/3, 1-\varepsilon$ corresponding to the 3 possible false types and the true type.

The mutation rate for different death-to-birth ratios 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.

Our implementation of the synthetic data generation method is available at `https://github.com/tg433/evolgenom`.

Calculating the likelihood

We calculate the average of the log-likelihood, Eq. (5), on a fixed grid of death-to-birth ratios, where for each point we generate a sample tree using the ELynx suite [28]. At each grid point the error rate is optimized either using Brent's method [29] implemented in Julia's Optim package [30] or calculating the log-likelihood on a grid of error rate values to speed up the calculation. The mutation rate is estimated in both cases according to Eq. (8). The final death-to-birth ratio is then estimated by interpolation using cubic splines and the final estimate for the mutation rate is based on this value.

Our implementation of the parameter estimation method is available at `https://github.com/tg433/evolgenom`.

Preprocessing of the empirical data

The raw sequencing data was preprocessed according to [15], using the code provided by the authors. The DNA contents of 10^4 cells were sequenced [16], along with a sample of neighbouring normal tissue. Mutations were called using VarScan 2 [31], which is flexible and easy to adapt to the requirements of the fitting procedure. The original study [15] targeted 410k sites, however, the raw data covers well over 10^8 sites. We applied the mapping quality and base quality thresholds of [15] to the data, and also left the minimum sequencing depth requirement of VarScan 2 at its default setting (8), both in the tumor and the normal samples. 923383 sites remained. Among these sites, 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 correspond better to the error model of our fitting procedure than an error rate which changes sharply with threshold frequency or read count 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 [16]. We also checked that the default somatic p-value threshold does not exclude any candidate somatic mutations. Other parameters were left at their default values. Mutation frequencies were corrected for *i)* copy number variation (CNV), using VarScan 2 with default parameters, and *ii)* 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 [32]. Sites having multiple variant types (i.e., number of reads of wild type plus most frequent mutant type being lower than the sequencing depth) were checked manually. Read counts of all 4 possible genotypes were identified for all variant sites.

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.

References

[1] Sottoriva A, Kang HK, Ma Z, Graham TA, Salomon MP, Zhao J et al. A Big Bang model of human colorectal tumor growth. Nat Gen. 2015;47:pages 209–216.

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

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

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

[5] 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.

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

[7] 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.

[8] Jones S, Chen W, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. PNAS. 2008;105:4283-4288.

[9] Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. PNAS. 2010;107:18545-18550.

[10] Sun S, Klenaber F, Tian T. A new model of time scheme for progression of colorectal cancer. BMC Sys Biol. 2014;8:S2.

[11] 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.

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

[13] Felsenstein J. Phylogenies From Molecular Sequences: Inference And Reliability. Ann Rev. 1988;22:521-565.

[14] 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.

[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] 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.

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

[18] Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nat Comm. 2017;8:15183.

[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] 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.

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

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

[27] 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.

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

[29] Brent RP. Algorithms for Minimization without Derivatives. Englewood Cliffs, NJ:Prentice-Hall; 1973.

[30] Mogensen PK, Riseth AN, White JM, Holy T. Optim.jl. Version 0.16.0 [software]. Available from: https://julianlsolvers.github.io/Optim.jl/stable/#.

[31] 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.

[32] 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.