Open access • Posted Content • DOI:10.1101/753871

Estimating growth patterns and driver effects in tumor evolution from individual samples — Source link

Leonidas Salichos, William Meyerson, Jonathan Warrell, Mark Gerstein

Institutions: Yale University

Published on: 02 Sep 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population and Mutation rate nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-14407-9 OPEN

Estimating growth patterns and driver effects in tumor evolution from individual samples

Leonidas Salichos[1,2], William Meyerson[1,2], Jonathan Warrell[1,2] & Mark Gerstein[1,2,3,4*]

Tumors accumulate thousands of mutations, and sequencing them has given rise to methods for finding cancer drivers via mutational recurrence. However, these methods require large cohorts and underperform for low recurrence. Recently, ultra-deep sequencing has enabled accurate measurement of VAFs (variant-allele frequencies) for mutations, allowing the determination of evolutionary trajectories. Here, based solely on the VAF spectrum for an individual sample, we report on a method that identifies drivers and quantifies tumor growth. Drivers introduce perturbations into the spectrum, and our method uses the frequency of hitchhiking mutations preceding a driver to measure this. As validation, we use simulation models and 993 tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium with previously identified drivers. Then we apply our method to an ultra-deep sequenced acute myeloid leukemia (AML) tumor and identify known cancer genes and additional driver candidates. In summary, our framework presents opportunities for personalized driver diagnosis using sequencing data from a single individual.

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [3] Department of Computer Science, Yale University, New Haven, CT 06520, USA. [4] Center for Biomedical Data Science, Yale University, New Haven, CT 06520, USA. *email: pi@gersteinlab.org

Over the past several decades, researchers have proposed different models to explain tumor progression, including stochastic progression, the mutator phenotype, and clonal evolution[1–3]. Originally suggested about 40 years ago[3], Navin and colleagues provided strong evidence that the "punctuated clonal evolution" model constitutes a major force in cancer progression. According to this model, tumor progression is an evolving system subject to selective pressure while accumulating thousands of mutations[4,5].

Advances in technology have allowed scientists to sequence thousands of genomes, revealing millions of variants per individual[6–8]. In cancer genomics, The Cancer Genome Atlas (TCGA)[4] offers access to thousands of cases encompassing over 30 types of cancer. Similarly, the International Cancer Genome Consortium (ICGC) recently announced "data release 26", which comprises data from more than 17,000 cancer donors and 21 tumor sites. Within ICGC, the pancancer analysis of whole genomes (PCAWG) study is an international collaboration to identify common patterns of mutations in over 2800 sequenced whole-cancer genomes[9]. As cancer databases continue to expand, the amount of fully sequenced genomes will continue to increase, with future plans setting goals for the storage of more than a million genomes[10]. Concurrently, deeper sequencing signifies less noise, more accurate variant-allele frequencies (VAFs), and more accurate subclonal and single-nucleotide variant (SNV) identification, while increasing the detection of novel drivers[11–13].

Recent studies have tackled the effect of selection in tumor progression in the context of clonal evolution, neutral evolution, and selection, providing valuable insights about the clonal progression of the disease[5,14–16]. By considering tumor progression as an evolutionary process, cancer development follows the trajectory of different evolutionary pathways based on cell and population dynamics, optimization strategies, and selective forces. These evolutionary trajectories have been shown to influence primary tumor growth[17] and the timing of landmark events[18]. However, the evolutionary and selective mechanisms during tumor progression remain unexplored and strongly debated[19–21].

Accumulated SNVs have been characterized as drivers or passengers, depending on whether or not they provide a selective advantage for the tumor cells. If the selective advantage or their respective effect is weak, the mutations are known as mini-drivers, although the existence and detectability of mini-drivers has been debated[22,23]. Identifying SNV and gene drivers has been one of the focal points of cancer genomics, where different methods aim to detect driver mutations based on selection, recurrence or changes in mutational density[22,24]. These methods rely on the deviation from our expectation of the underlying genomic mutation rates, often by considering additional covariates such as replication timing and gene expression[25–27]. Other methods, characterized as ratiometric, assess the composition of mutations, normalized by the total mutations in a gene[22]. This includes the proportion of inactivating mutations, recurrent missense mutations, functional impact bias, mutational composition, or clustering patterns[28–31]. However, if only a small proportion of mutations within a genomic region (which is potentially under negative selection or functional restrictions) facilitates cancer progression, driver detection requires either a very large sample, a strong effect or otherwise the driver's presence is undetectable[22]. Further, mutational heterogeneity in cancer poses an additional problem for large cohorts; as the sample size increases, so does the list of putatively significant genes, producing many false-positive driver genes[26]. More importantly, only a minimal portion of driver mutations are, in fact, true drivers[32]. This is particularly important in a clinical context as assessing a cancer gene mutation as a true functional driver is a critical problem for drug selection[32,33].

According to recent studies[34] and in agreement with past theories[35], a few major genetic hits (strong drivers) can induce tumorigenesis. At the same time, a driver mutation may not actually be the cause of tumorigenesis, but instead only increase growth rate and therefore be under positive selection[36]. One of the most common and widely used lists of cancer genes is the Vogelstein list[28], consisting of ~140 oncogenes and tumor-suppressor genes (TSGs). While high-impact mutations in TSGs might favor cancer progression by deactivating tumor suppression, oncogenes need altered expression levels to favor tumor growth. Thus, high-impact mutations, such as nonsense mutations in oncogenes might decrease gene expression and burden tumor cells[37]. Less appreciated is the role of noncoding mutations in tumor progression[36,38,39]. Interestingly, in the case of TSGs, different studies have reported the role of noncoding intronic mutations that alter correct exon splicing, resulting in faulty tumor suppression[40–43]. Similarly, in the case of oncogenes different studies have reported the potential effect of synonymous mutations[39,40,44]. For example, Gartner et al.[44] showed that the early synonymous mutation F17F in the *BLC2*-like 12 gene alters the binding affinity of regulatory hsa-miR-671-5p, leading to changes in expression.

In our study, we developed a framework to model tumor progression and the effect of drivers in individual deep-sequenced tumors. We successfully applied our model using 993 linear tumors (linear subclonal expansion, where each parent subclone has one child subclone) from the PCAWG consortium, and found that predicted drivers[9] are associated with periods of positive growth. Our results suggest that mutations involved in biological processes such as cell development, cell differentiation, and multicellularity appear under strong positive or negative growth enrichment. Missense or nonsense mutations in TSGs were enriched during positive growth. We also identified significant positive enrichment for mutations in the promoter regions of both TSGs and oncogenes. In addition, in the case of TSGs, we discovered a small but significant signal from intronic mutations. Finally, we applied our framework to a deep-sequenced model acute myeloid leukemia (AML) tumor, where our predicted growth peaks aligned closely with three missense mutations from known cancer genes. Notably, our analysis suggests the potential presence of additional driver candidates.

Results

Method formalism. When sequencing a cell population or tumor bulk, each mutation is assigned a variant-allele frequency (VAF), which corresponds to the mutation's frequency in the resulting pool. According to the infinite sites model[45], once a mutation occurs it will continue to exist within that cell and its descendants. Therefore, if we assume that there is no selection or chromosomal duplications, the VAF is associated with the time of occurrence and population growth rates. That is, in the presence of a driver (i.e., in cells with higher fitness), nondriver mutations within that cell lineage will also have higher-than-expected VAF and are termed "hitchhikers"[28] (Fig. 1 and Supplementary Notes). Hitchhikers that initially occurred before the driver mutation but continue to exist within that cell lineage will have a VAF that is higher than or equal to the driver's frequency. We call these hitchhikers "generational" (g-hitchhikers) because they essentially mark the different generations of an ever-increasing number of tumor cells and thus exhibit a clock-like behavior. Since any nondriver lineage derived from the division of earlier cells will result in a mutation having lower frequency, these predriver hitchhiking mutations will indicate generational growth (Fig. 1). As the fitness mutation becomes more prevalent over time, so

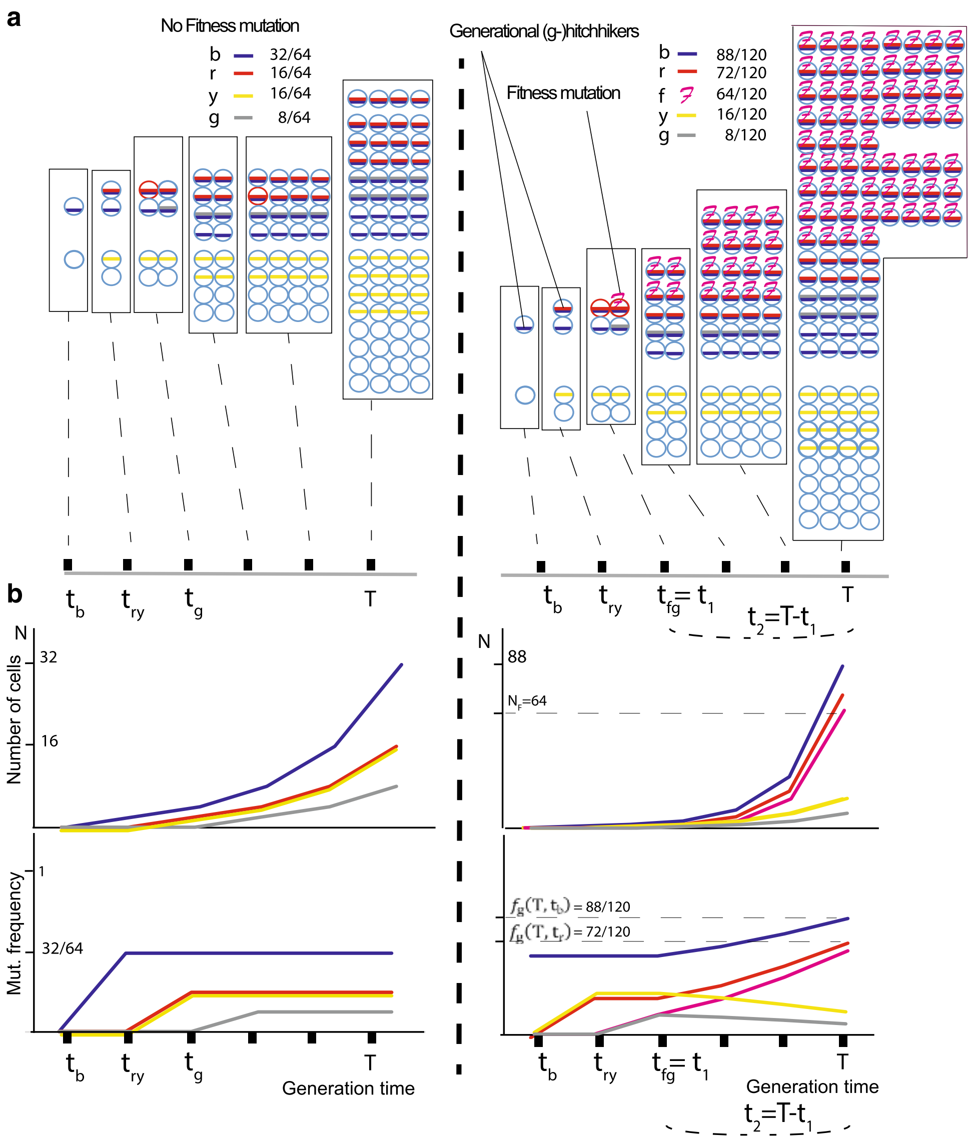


Fig. 1 (g−)Hitchhikers' frequency depends on driver's effect. We consider a simple population of cancer cells that grows exponentially $N(t) = \mathrm{e}^{rt}$; for simplicity, we assign one mutation per cell division. At the time of biopsy T, the frequency of a mutation occurring at time t_n would be equal to $f_{\mathrm{n}}(T, t_{\mathrm{n}}) = \frac{\mathrm{e}^{r(T-t_{\mathrm{n}})}}{\mathrm{e}^{rT}} = \mathrm{e}^{-rt_{\mathrm{n}}}$. At time t_1, a mutation occurs that increases the growth rate r of the specific subpopulation by a scalar multiplier k, such that the new population is now expanding as $N_{\mathrm{F}} = e^{krt_2}$. Thus, at the time of biopsy $T = t_1 + t_2$, we expect a generational (g−) hitchhiking mutation that occurred at time $t_m < t_1$ to have a frequency equal to $f_{\mathrm{g}}(T, t_{\mathrm{m}}) = \frac{\mathrm{e}^{r(T-t_{\mathrm{m}})} + N_{\mathrm{F}} - \mathrm{e}^{rt_2}}{N_{\mathrm{tot}}}$, where N_{tot} is the total number of cells (or mutations) and N_{F} is the number of cells that contain the fitness mutation that occurred at t_1 and expanded for t_2. Therefore $N_{\mathrm{F}} = e^{krt_2}$. In **a**, we show the mutational frequencies at the time of biopsy T for two growth models; one neutral and one with a fitness mutation occurring at time $t_1 = t_{\mathrm{fg}}$. Hitchhiking mutations "b" (blue), "r" (red), as well as passenger mutations "g" (gray) and "y" (yellow), also occur at different time points. **b** Under an exponential model with a fitness mutation occurring at time $t_1 = t_{\mathrm{fg}}$, hitchhikers "b" and "r" show an increased frequency compared to neutral, subject to time and effect of the fitness mutation. Passenger mutations "y" and "g" that occurred before or with the fitness mutation, but on a different cell lineage, end up with lower frequencies. We characterize mutations "b" and "r" as generational ($g-$) hitchhikers since they mark the population's generational growth.

does the prevalence of predriver g-hitchhikers, but critically at a different pace, which we calculate (see Methods).

Our framework's equations relate the VAF of generational hitchhiker mutations to the fitness effect of the subclonal driver with which they are hitchhiking, mediated by various growth and population parameters (i.e., the base growth rate r, a scalar multiplier k corresponding to fitness effect of the mutation, the time t_1 when the driver mutation is generated, N_{tot} the population size and N_{F} the driver's subclone size). The existence and fitness effects of subclonal drivers are not directly observable but are of primary biomedical importance. The VAF of hitchhiker mutations is directly observable, therefore we chose to use these VAFs to infer the presence of subclonal drivers and estimate their fitness effects. Our approach is to fit the known VAFs of the hitchhiker mutations in the hitchhiker equations to estimate the growth pattern and the fitness effect of subclonal drivers. This method requires to simultaneous estimate the various growth and population parameters, which we performed using nonlinear least-squares optimization. To address the fact that real tumors differ from idealized behavior, we make use of sliding windows and local timepoint reoptimizations in the parameter estimation to prevent departures from idealized behavior in one part of the VAF spectrum from interfering with parameter estimation in other parts of the VAF spectrum. We derived our estimators for r and k through the implementation of a deterministic model to a stochastic process with a large final population N_{tot}.

Modeling the frequency of g-hitchhikers. We assume a simple and neutral population of cancer cells that grows exponentially with rate r. For simplicity, we here assign each new daughter cell one new mutation (alternative mutation rates do not affect the derivation, see Methods). At time t_1, a mutation occurs that accelerates the growth rate of the specific subpopulation by a scalar multiplier k such that the new population expands with new rate $k \times r$. At the time of biopsy $T = t_1 + t_2$, where the fitness mutation occurs at t_1 and expands for time t_2, we expect the frequency of a generational g-hitchhiker mutation that occurred at time $t_{\mathrm{m}} < t_1$ (see Fig. 1 and Methods) to follow a frequency function f_{g}

$$f_{\mathrm{g}}(T, t_{\mathrm{m}}) = \frac{N_{\mathrm{R}} + N_{\mathrm{F}} - N_{\mathrm{RF}}}{N_{\mathrm{tot}}},$$

or

$$f_{\mathrm{g}}(T, t_{\mathrm{m}}) = \frac{e^{-rt_m}\left[N_{\mathrm{tot}} - f_{d(T,t_1)} \times N_{\mathrm{tot}} + \sqrt[k]{f_{\mathrm{d}(T,t_1)} \times N_{\mathrm{tot}}}\right] + f_{\mathrm{d}(T,t_1)} \times N_{\mathrm{tot}} - \sqrt[k]{f_{\mathrm{d}(T,t_1)} \times N_{\mathrm{tot}}}}{N_{\mathrm{tot}}}, \quad (1)$$

where $f_{\mathrm{d}(T,t_1)}$ is the frequency of the driver mutation occurring at t_1 and expanding for $t_2 = T - t_1$, The terms $\left\{\mathrm{e}^{-rt_{\mathrm{m}}} * \left[N_{\mathrm{tot}} - f_{\mathrm{d}(T,t_1)} * N_{\mathrm{tot}}\right]\right\}$ and $\{f_{\mathrm{d}(T,t_1)} * N_{\mathrm{tot}}\}$ correspond to the growth of regular N_{R} and fitness N_{F} populations respectively, while extracting $N_{\mathrm{RF}} = \left\{\sqrt[k]{f_{\mathrm{d}(T,t_1)} * N_{\mathrm{tot}}}\right\}$ for not double counting the hypothetical regular growth of fitness cells.

Eq. (1) for the mth hitchhiker implicitly allows one to use the previous $m - 1$ potential hitchhikers to refine the estimates of growth rate r and scalar effect k. This estimation is achieved either through a nonlinear-least-squares optimization, and/or through the independent calculation of growth r.

The frequency of g-hitchhiking mutations follows the form of an exponential distribution. Theoretically, this further allows us to estimate growth rate r from consecutive g-hitchhiking mutations m_1, m_2, and m_3, which occurred at times $t_{\mathrm{m}1}$, $t_{\mathrm{m}2}$, and $t_{\mathrm{m}3}$ ($t_{\mathrm{m}1}$, $t_{\mathrm{m}2}$, and $t_{\mathrm{m}3} < t_1$), respectively, according to

$$r = \ln\left(\frac{f_{\mathrm{g}}(T, t_{\mathrm{m}1}) - f_{\mathrm{g}}(T, t_{\mathrm{m}2})}{f_{\mathrm{g}}(T, t_{\mathrm{m}2}) - f_{\mathrm{g}}(T, t_{\mathrm{m}3})}\right). \quad (2)$$

In practice, to obtain more accurate estimates, our default algorithm estimates the growth rate r from three more distant time points t, $t + n$, and $t + m$ ($n < m$ and $t + m < t_1$) with final frequencies $f_{\mathrm{g}}(T,\ t)$, $f_{\mathrm{g}}(T,\ t_{\mathrm{n}})$, and $f_{\mathrm{g}}(T,\ t_{\mathrm{m}})$, respectively (see Methods).

Optimizing for any time point during tumor progression. In addition to our independent estimate of growth rate r, and in order to avoid previous frequency perturbations in our sample and localize the effect timewise, we also include an extra parameter referred to as "generational time (t_{g})", which allows us to calibrate an offset for the number of past generations until that point without considering previous mutations outside our sliding window. Thus, similar to Eq. (1), we now have

$$f_{\mathrm{g}}\left(T, t_{\mathrm{g}}, t_{\mathrm{i}} - m\right) = \frac{\mathrm{e}^{-r\left(t_{\mathrm{g}} + t_{\mathrm{i}} - m\right)} \times \left(N_{\mathrm{tot}} - f_{\mathrm{d}(T,t_i)} \times N_{\mathrm{tot}} + \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}}}\right) + f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}} - \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}}}}{N_{\mathrm{tot}}}, \quad (3)$$

where $f_{\mathrm{d}}(T, t_{\mathrm{i}})$ is the frequency of the putative driver i occurring at time t_{i}.

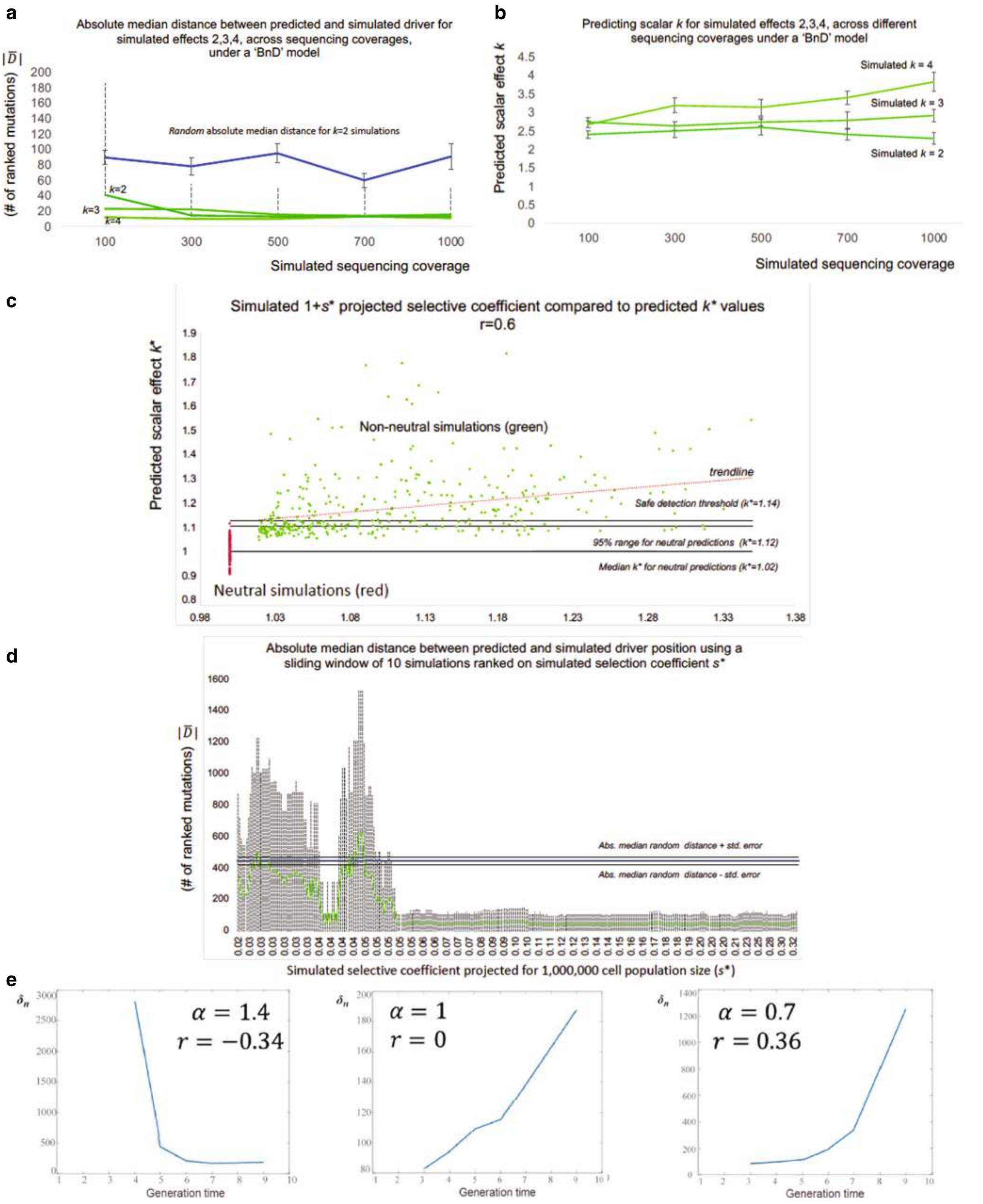
This approach allows us to reoptimize t_{g} at any time t_{i} during tumor growth, independently of earlier or later calculations.

Birth-and-death model, Gillespie simulations. First, we tested our algorithm on simulated data based on various growth models, including (a) exponential growth, (b) exponential growth with delayed cell division, and (c) logistic growth (birth-and-death model). We performed simulation models (a) and (c) using a stochastic Gillespie algorithm, whereas model (b) represents an exponential cell growth model with a lag time for cell division, which prevents a cell from redividing immediately. Briefly, for the Birth-and-Death Gillespie model, which is the workhorse of our simulations, we used a stepwise time-branching process to model the growth of a single transformed cell into a tumor with a dominant subclone. At each time step, an event type is chosen with a probability proportional to the event's prevalence. Then, a cell of the eligible type is randomly chosen to undergo that event. In our logistic-growth simulations, the death rate of each cell climbs proportionally as carrying capacity is reached, whereas in our exponential simulations, the death rate of each cell is constant throughout the simulation. The simulation ends randomly, after the driver subclone reaches a critical prevalence. The Gillespie algorithm has been frequently used to simulate stochastically dividing cells[46–52], although simulations with special attention to cell cycle have also been recommended[53].

During simulated growth, we assigned a driver mutation with additional propagating effects from nearly neutral to high ($k =$ 1.1, 2, 3, and 4), thus leading to faster growth for the respective subpopulation that contains the specific mutation. Using conservative assumptions, these scalar values represent a range of projected selection coefficients s^* from 0.001 to 0.03 in biologically sized populations (see Methods). For each simulation, we calculated each mutation's frequency in the total population and ordered them based on that frequency. Then, by applying our method we calculated the ranking distance D (as the number of ordered mutations) between the true and our predicted driver (growth peak), as well as the driver's scalar effect k.

We tested our method's performance in simulated tumors of lower coverage and different effects. Higher sequencing depth and scalar effect k provided more accurate results and improved our method's implementation (Fig. 2a, b). Lower coverage was associated with worse k calculations and driver predictions, as well as lower positive-predictive values (PPVs). For weak drivers, low sequencing coverage made their identification more difficult. Absolute median ranking distance $|\widetilde{D}|$ was 41 for coverage 100 ×/$k = 2$, compared to 13 for coverage 1000×/$k = 2$ and $|\widetilde{D}| = 11$ for coverage 1000×/$k = 4$, respectively. In general, driver identification required either a higher than 100× coverage, or a stronger effect (i.e., $k > 2$, $s^* > 0.01$ for a projected cell population of 1,000,000 cells) (Fig. 2i).

Overall, we were able to well approximate the driver's occurrence and effect (Fig. 2). For the birth-and-death model with simulated coverage 1000×, the median predicted estimation for simulated effects $k = 2$, $k = 3$, and $k = 4$ was 2.3, 2.9, and 3.8, respectively (Fig. 2ii). Moreover, the median ranking distance $\widetilde{D}$ between simulated and predicted drivers with effect $k = 1.1$ (nearly neutral), $k = 2$, and $k = 3$ was 71, 3. 5, and 6, respectively. The corresponding median distances for random mutations were 73, 43, and 41 (Supplementary Fig. 1). For our nearly neutral simulations ($k = 1.1$, s^* ~0.001 for a projected cell population of 1,000,000 cells) the median distance $\widetilde{D}$ in driver predictions and random predictions was very similar and not significant.



Simulations with added stochasticity in mutation rates. To further test our model on a separate independent simulation dataset, we applied our method to (a) neutral simulations of tumor progression and (b) nonneutral simulations for various growth scenarios, as previously developed and described by Williams et al.[16,54] (see also Supplementary Methods). These simulations, although also based on the Gillespie growth model, included added stochasticity with varying mutation rates during tumor progression ($\bar{\mu} = 10$ mutations per cell division). For every simulation, both neutral and nonneutral, we identified our model's highest predicted effect peak, calculated the effect k and absolute median ranking distance $|\widetilde{D}|$ between the simulated and

Fig. 2 Deeper coverage and stronger drivers improve predictions. In **a**, using 541 simulations of tumor growth under a birth-and death model, we show the absolute median distance $\widetilde{|D|}$ as in "absolute number of ordered mutations" between predicted and simulated driver for sequencing depths. With the exception of $k = 2$ for 100× (two-tailed t test $P = 0.015$), we were able to detect the driver's presence ($P < 0.005$). Blue line represents the random $\widetilde{|D|}$ as derived by selecting a random mutation from each simulation and calculate the absolute distance to the simulated driver. Dotted lines represent the $2 \times \sigma$ deviation from $\widetilde{|D|}$ while capped bars the median's standard error. For convenience, we only show bars for $k = 2$. In **b**, Using the same simulations, lower coverage provides less accurate k predictions with a lower effect. Capped bars represent the standard error of the median effect prediction. The three lines represent simulations with simulated effect of 2–4. In **c**, using the "Williams et al. 2018" algorithm, we simulated 360 nonneutral and 140 neutral tumors for 10,000 cells. Then, we adjusted our effect predictions for n^* equal to 1,000,000. In addition, we also adjusted the simulated selection coefficient s^* for the same populations. Pearson's r between the simulated adjusted coefficient "$1 + s^*$" against adjusted predicted k^* was 0.6. In **d**, after ranking s^* for every nonneutral simulation, we used a sliding window of 20 simulations to estimate $\widetilde{|D|}$ (and $2 \boldsymbol{\times} \sigma$) between the simulated and predicted driver within every window. Dotted lines represent $2 \times \sigma$ deviation. When $s^* > 0.05$ our driver detection became highly accurate. Blue line represents $\widetilde{|D|}$ for random predictions (444.5), while black lines represent median standard error (24.5). Simulated s^* have been projected for $n^* = 1{,}000{,}000$. In **e**, using Kingman's coalescent theory, we show that growth estimator $\hat{\boldsymbol{r}}$ remains qualitatively unchanged even for non g-hitchhikers. As mutational density $\delta_{\rm n}$ increases with n, and hence with time, $\hat{\boldsymbol{r}}$ estimator is predicted to take positive values for both constant and varying populations. Similarly, for negative growth, $\boldsymbol{\delta}_{\rm n}$ decreases with time. We let $\alpha > 1$ corresponding to a decreasing and $\alpha < 1$ corresponding to an increasing population.

predicted driver in number of ranked mutations. Various scenarios for nonneutral growth included a wide range of simulated selection coefficients s (0–33, for a population size of 10,000 cells), categorized driver's VAF (small 0.1–0.2; medium 0.2–0.3; large 0.3–0.4) and larger cell population projections using population genetic models and method adjustments. Corresponding neutral simulations were also generated using the same population parameters. Overall, and in agreement with our previous analyses, our results suggest a small overlap between neutral and nonneutral peaks for weak drivers (Fig. 2c and Supplementary Fig. 1f) and highly significant driver predictability when the predicted driver effect was larger than our (narrow) neutral-effect distribution (Fig. 2c, d and Supplementary Fig. 1g–i). For instance, for simulated populations of 10,000 cell without projection ($0 <$ simulated $s < 33$) and 1000× coverage our method provided accurate driver detections when the predicted effect was larger than $k = 1.29$ with $\widetilde{|D|}$ ~50 mutations compared to 444.5 for random. These results are directly comparable to our previous analyses, considering the new mutation rates. Similarly, for a projected cell population of 1,000,000 cells, our method provided accurate driver detection for projected selection coefficient $s^* >$ 0.05 (Fig. 2d). Larger population projections typically decreased the predicted effect k^* and selection coefficient s^*, but did not affect our method's ability to detect drivers (Supplementary Fig. 1k) as these projections also decreased the standard deviation of our neutral-effect distribution (predicted k^* for neutral-effect peaks). When we combined 140 neutral with 360 nonneutral simulations, drivers with medium final VAF showed the highest correlation between simulated selection coefficients and our method's predicted scalar k effects (Pearson correlation $r = 0.60$, Fig. 2c). Drivers with lower final VAFs (small ~0.1–0.2) provided slightly lower correlation but had the highest driver detectability, with $\widetilde{|D|} = 46$ mutations between the simulated and predicted driver (Supplementary Fig. 1l), where random $\widetilde{|D|}$ was 444.5 mutations. A (tenfold) higher $\widetilde{|D|}$ here is expected since for these simulations we assumed 10 instead of 1 mutation per cell division.

Estimator $\hat{r}$ for non g-hitchikers. We also tested the behavior of the estimator for r (Eq. (2)) on non-g-hitchhiking mutations (i.e., when the assumption that the mutations are generational hitchhikers is not satisfied). For this purpose, we used coalescent theory to estimate the variation in density of mutations across the VAF spectrum for a variety of models (see Methods). We first analyzed the behavior in a constant-size population, and then in populations with increasing and decreasing exponential growth. Our analysis shows that the growth indicator does not qualitatively change its behavior in this context, so that negative values continue to represent periods of negative growth, and large positive values represent periods of positive growth. However, here we expect a small positive value in the case of zero growth (Fig. 2e, Supplementary Fig. 2).

Growth patterns in 993 linear tumors from PCAWG. Using 993 linear tumors from the PCAWG consortium, we explored the different patterns and dynamics of tumor growth based on our model's assigned growth rates. Tumor linearity (where no parental subclone has two or more children subclones) further ensures that tumor subclones do not intermingle and that higher VAF is associated with earlier occurrence. We note that mutational frequency as described in our equations corresponds to 2× VAF, with correction for purity and copy-number variations. These VAF corrections were obtained from PCAWG and are not implemented by our method, which only considers a final mutational frequency. Using our model, each mutation i from sample in our database is assigned a potential positive or negative growth value $r_{\rm i}$ and a driver effect $k_{\rm i}$. Under ideal conditions, for each sample, a vector of effect peaks $r_{\rm i-1} \times k_{\rm i}$ corresponds to potential drivers at position i. However, noise, coverage, and growth stochasticity can cause these peaks to represent the potential presence of a nearby driver, especially in low-coverage sequenced tumors (see Fig. 3a, b).

To identify growth patterns across individual tumors, we (i) normalized each mutation's growth rate based on the sample's maximum growth value; (ii) divided the ordered mutations into 20 bins; and (iii) applied k-means clustering to the average normalized value per bin. Our results highlighted three main clustering patterns (Fig. 3c). As expected, most tumors ($n = 525$) showed logistic growth with an increasingly higher growth rate at the beginning and a stabilization at the later stages. For many tumors ($n = 366$), an early high growth period was followed by a stagnation and potential reduction in tumor size. This effect could also be artificially enhanced due to sampling errors for mutations with low VAF (during late tumor progression). The last group of tumors ($n = 102$) showed relatively steady, continuous growth. However, it is uncertain whether this pattern represents tumors that were sequenced early. Further, some types of cancer seemed to prefer specific growth patterns (Fig. 3c).

By modeling tumor growth, we can find mutations during positive or negative growth periods in single or multiple individual samples. Through positive growth enrichment, we characterized the degree to which one type of mutation (e.g., TSGs/*TP53*, nonsynonymous) or region (e.g., *TP53*) was significantly enriched and associated with periods of positive growth across multiple samples. We then compared each

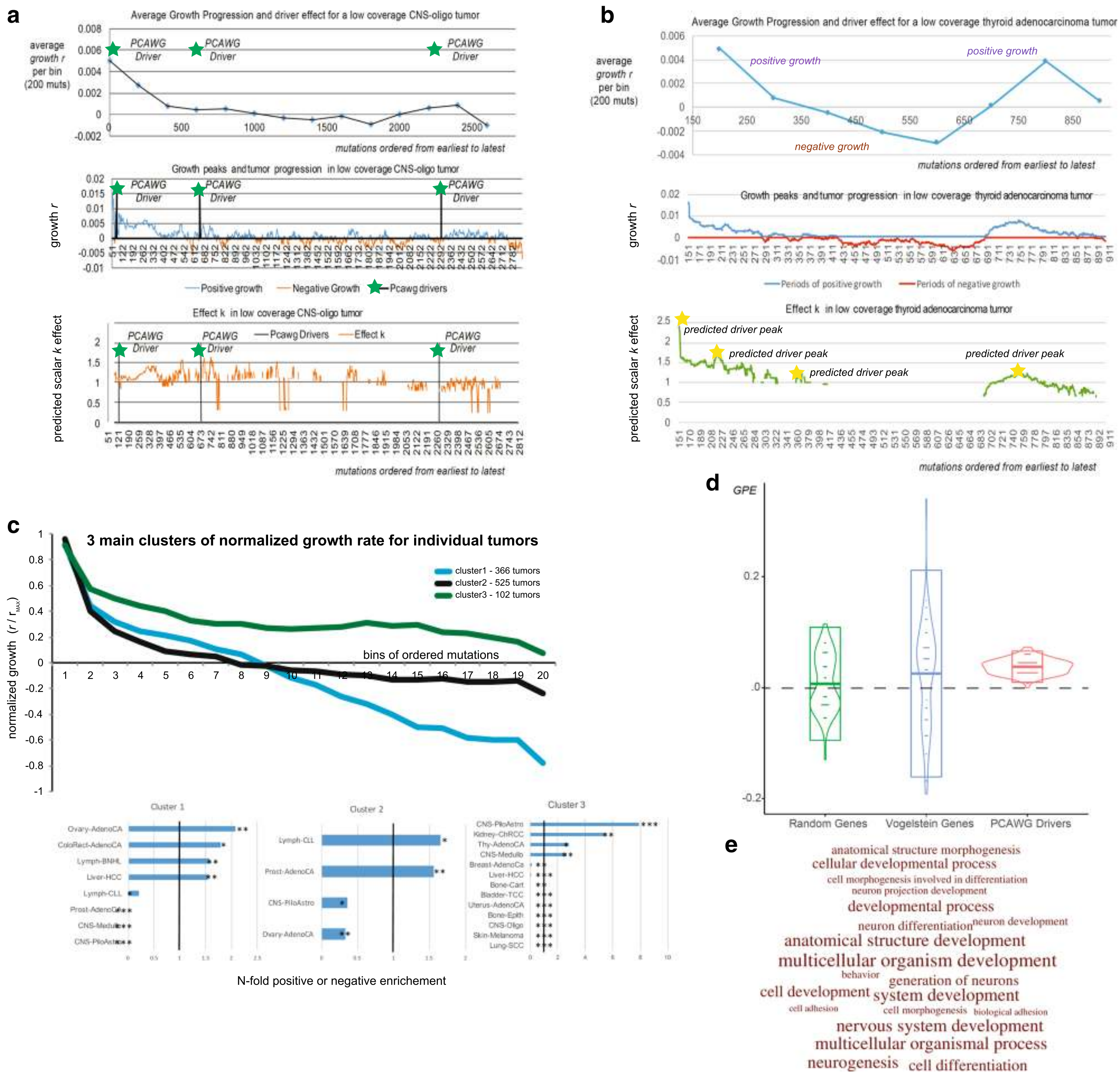


Fig. 3 Growth patterns and growth associations using 993 linear tumors. Across 993 linear tumors from PCAWG consortium we expect an under-selection mutation to be associated with periods of positive growth. We compared several mutation types (driver mutation, mutation within geneX, within GO categoryX), to a random distribution from their respective sample for association with positive growth. **a**, **b** The averaged growth progression, mutational growth, and mutational effect, for a single low-coverage CNS-oligo tumor and a single low-coverage thyroid adenocarcinoma tumor without any PCAWG-identified drivers. Green asterisks denote the ordered position of a PCAWG-predicted driver within the sample. Yellow asterisks denote a growth peak and putative driver presence. In **c**, we derived three main growth patterns (steady growth, sigmoid growth, and stagnation/shrinkage) for 993 linear tumors, as they were grouped using a *k*-means clustering algorithm. Various cancer types showed specific enrichment or depletion for the three clusters (levels of significance for Fisher's tests for enrichment noted as *, **, and *** for $p < 0.05$, 0.01, and 0.001). In **d**, PCAWG drivers and Vogelstein genes show significant positive growth enrichment compared to a list of random highly mutated genes. Boxplots represent $2 \times \sigma$ deviation, lines represent the mean, while violin plots are trimmed to data range. **e** We show the GO enrichment for the 20 most affected biological processes, when we use 293 genes, significantly associated with periods of positive growth.

mutation type to random mutations from their respective samples. To confirm whether we could detect any signal of selection at the gene level, we compared positive growth enrichment (PGE) for mutations between (i) the Vogelstein gene list[28]; (ii) a comparable list (in mutational numbers) of randomly selected genes; and (iii) a list of assigned drivers from the PCAWG consortium[9,32]. As expected, PCAWG-assigned driver SNVs clearly showed the highest positive enrichment, followed by SNVs that were not individually called by PCAWG as drivers but that fall within the Vogelstein driver gene list (Fig. 3d). We note, however, that our random gene list did show a small positive enrichment, as this list contains several often-mutated genes and potential drivers or mini-drivers. We obtained similar results when we repeated the comparison while considering the difference between additional mutational effect against a random distribution (Supplementary Fig. 3).

In an effort to better understand the microenvironment of tumor dynamics, the selective forces, and the biological processes that are most keenly affected by tumor progression, we analyzed a list of 1000 most mutated genes in the PCAWG samples where we identified 293 genes with significant overall association with positive growth (Supplementary Data 1). Then we further tested these genes for Gene Ontology (GO) enrichment. As expected, developmental and differentiation processes were highly enriched during periods of positive growth, showing signals for being under positive selection. Interestingly, we found that genes related to multicellular processes showed the highest enrichment based on raw p value (Fig. 3e, Supplementary Data 2).

TSGs vs. oncogenes. Based on each mutation's genomic properties (e.g., genomic position, coding vs. noncoding, TSG vs. oncogene, cancer type, and GO annotation), we can examine whether the specific type of mutation (or mutation element) is statistically enriched during periods of positive growth when compared to random mutations from their respective samples. However, the more specifically that we defined a mutation type, the fewer mutations that corresponded to this category. For example, the Vogelstein TSGs in our dataset contain 321 missense and 103 nonsense mutations, whereas *TP53* in our dataset contains 71 nonsynonymous mutations and 13 nonsense mutations. Unfortunately, for many tumor genes and cancer types, we currently have a small number of mutations, precluding significance in the results.

A recent study by Kumar et al. suggested that high-impact mutations should have more clear positive effects on tumor growth when they are located in TSGs vs. oncogenes[37]. This is expected, as generally a "defected" oncogene with reduced expression should not favor cancer progression. To better understand the behavior of TSGs and oncogenes, we tested for positive enrichment of synonymous, nonsynonymous, premature stop, promoter, and intronic mutations (Fig. 4). As expected, our results showed significant enrichment of missense and nonsense mutations in TSG regions. During periods of positive growth, 45 nonsense and 128 missense mutations corresponded to an average of 37.4 and 117.96 random mutations, respectively (100 bootstraps replicates, p values = 7.823348e−30 and 1.632649e−23). Interestingly, promoter and intronic regions also showed a significant positive effect on tumor growth, suggesting that some noncoding mutations in TSGs might favor positive growth (Fig. 4a).

In the case of oncogenes, we did not find significant enrichment of missense mutations, but we did find significant association between their promoter regions and positive growth (Fig. 4b). This might be due to many reasons including the pancancer nature of our analysis, lack of power and small sample size, our modeling assumptions, or the noise due to low sequencing coverage per tumor sample. However, many genes including oncogenes might be under negative selection, with only a small subset of their respective mutations being favorable to cancer growth. Moreover, high-impact mutations in oncogenic regions do not necessarily favor tumor growth. Indeed, our data contain only four nonsense mutations in oncogenic regions. Some oncogenes such as *MET* and *CTNNB1* showed slight overall negative enrichment, but their nonsynonymous mutations, especially in specific cancers, showed enrichment during periods of positive growth (Supplementary Fig. 3).

To detect mutations during positive growth periods, we applied our model to individual types of mutations (i.e., missense, synonymous, intronic, nonsense, and promoter) for each Vogelstein gene. Overall, our results identified various mutation elements including promoters, nonsense, and missense with significant effects (Fig. 4c). Interestingly, synonymous *BLC2* mutations that occurred near an early positioned mutational hotspot were significantly associated with positive growth (Fig. 4c and Supplementary Fig. 4). Synonymous mutations are not generally considered to be important in cancer; however, previous studies have reported recurrent synonymous F17F mutations in *BLC2*-like 12, where regulatory hsa-miR-671-5p alters the gene's expression[44].

Growth peaks and driver effects on a model AML tumor. In addition to the 993 PCAWG low-coverage tumor samples, we implemented our model on an ultra-deeply sequenced AML (>250×) liquid tumor. A ultra-deeply sequenced tumor provides more accurate global VAF, which should in turn allow for better estimation of model parameters[12].

In general, the predicted peaks of our model mapped very closely to mutations from known cancer genes (Fig. 5). Deep valleys followed by the highest growth peaks corresponded with close approximation to the three missense mutations from known cancer genes (*IDH1*, *IDH2*, and *FLT3*, p value < 2.2e−16). Thus, in agreement with previous studies[34,35], the derived growth patterns suggested three to five major genetic hits from cancer mutations in order to render tumor growth permanent.

In addition, we used all the mutations in our previous database to evaluate those in the deeply sequenced AML in order to identify new candidates associated with positive growth. As a result, we further identified five additional candidates from the ultra-deep AML sample that belong to genomic elements associated with positive growth (Fig. 5d). These additional candidates consist of four missense mutations (*SRCAP*, *CPS1*, *GLI1*, and *COL18A1*) and one intronic mutation (*MAP3K1*), which appeared to align near observed, previously unexplained periods of initial growth. Previous recent studies have also linked *CPS1* and *GLI1* to various cancers[55–58]. Finally, based on our PCAWG database, for each driver candidate we detected possible positive enrichment across varying effect ranges [0.9, 1.1, 1.3, 1.5, 1.7, 1.9, and 2.1] (Supplementary Fig. 5). Indicatively, our independent estimation of mutational effect suggested a high correlation when compared to the calculated effect using the deep-sequenced model AML tumor (Supplementary Fig. 5).

Discussion

Most approaches to identify driver candidates are based on recurrent mutations and large cohorts[22]. More recently, studies have probed tumor selection either through deviation from background metrics or by using VAF distribution to quantify the subclonal effect[20]. Here, we present a framework that models tumor progression using generational hitchhikers and localized time reoptimizations using mutational frequencies from individual samples to (i) determine periods of positive or negative growth, (ii) suggest the presence of candidate drivers and estimate their effect on tumor progression, and (iii) detect genomic regions or mutation elements that are associated with positive or negative growth periods. Overall, our work highlights the importance of whole genome deep sequencing for modeling tumor progression.

When we applied our framework to 993 individual tumors from the PCAWG consortium, our growth analysis indicated different growth patterns across cancer types, including steady growth, sigmoidal growth, and modes of stagnation. Determining tumor progression can be useful in understanding each tumor's historic aggressiveness, and the effect of driver mutations on tumor progression (VAFs used by our method typically represent past growth, as latest mutations tend to have undetected frequency in our sample). In addition, we identified several biological processes significantly affected by tumor progression, including genes involved in multicellularity. These results might

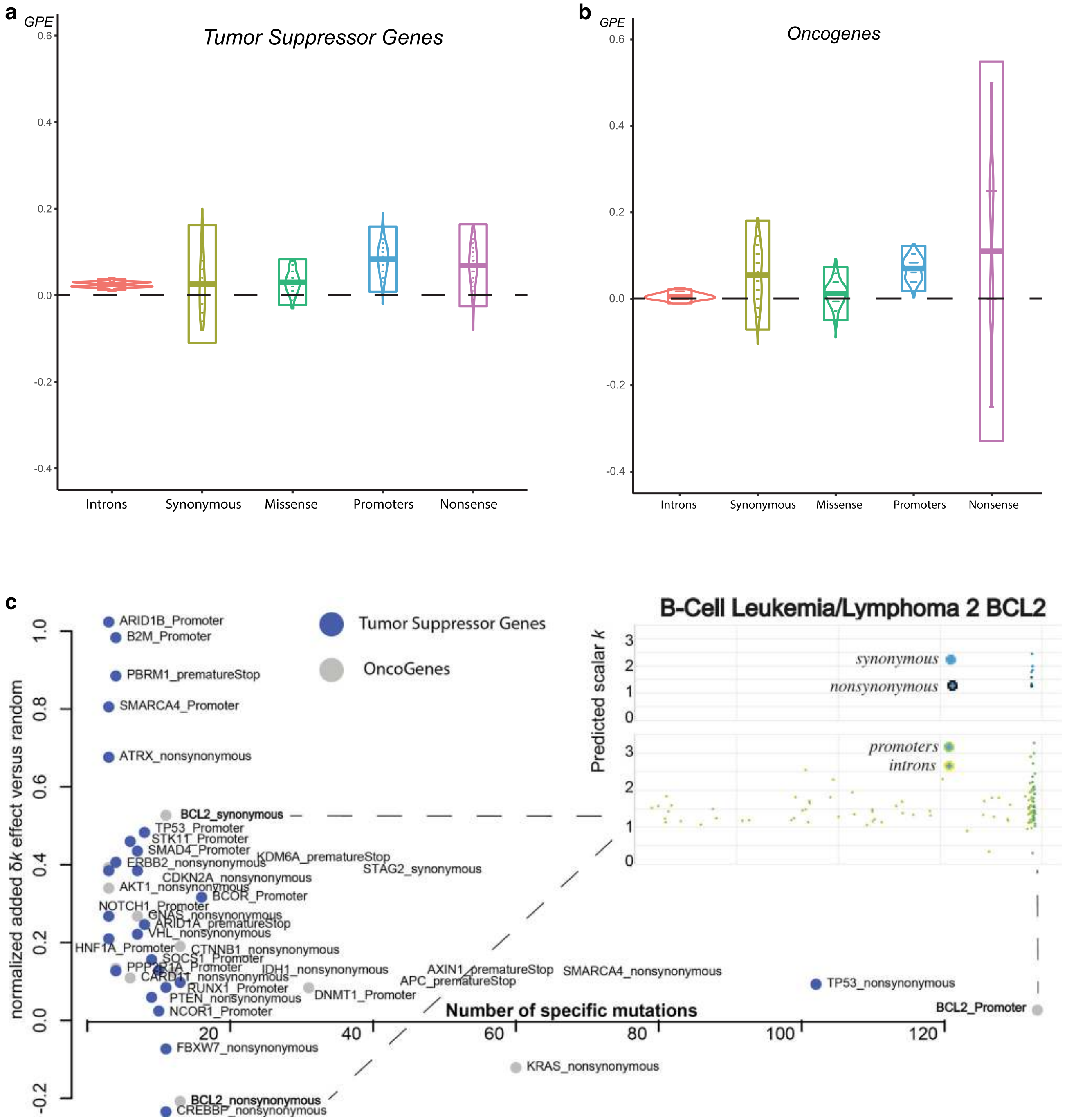


Fig. 4 Tumor-suppressor gene and oncogene elements show growth enrichment. We show the positive growth enrichment across different mutation types (introns, synonymous, missense, nonsense, and promoters). For **a**, Vogelstein tumor suppressor genes and **b** Vogelstein oncogenes, boxplots represent 2 × σ deviation, lines represent the mean, while violin plots are trimmed to data range. In **c**, we plot gene elements (e.g., {GeneX_mutation type}) from Vogelstein gene list that showed significant positive or negative enrichment. We further zoom in to *BCL2*'s genomic region to map missense, nonsynonymous, promoter, and intronic mutations.

indicate an evolutionary transition during tumor progression from multi-cell functionality to single-cell selection.

As expected, we found significant enrichment of known PCAWG drivers, Vogelstein cancer genes, and nonsense and missense mutation TSGs during periods of positive growth. In accordance with some previous studies[40–43], our results also suggested that a small proportion of intronic mutations could affect TSGs (but not oncogenes), whereas some synonymous mutations could affect oncogene (but not TSG) expression. Even though defective splicing in TSGs or changes in the negative regulation of oncogenes are not entirely unexpected[44], noncoding mutations are not generally considered to be major driver events in tumor progression. Thus, it is possible that our results are subject to analytical (e.g., model parametrization, initial parameters, window size selection, low sequencing coverage, and sample size) and biological (e.g., hitchhiking) error.

Using variant-allele frequency to quantify driver effects and tumor progression can be challenging. Our analysis might be subject to different types of bias, including sequencing noise, growth stochasticity, model parameterization, low sequencing

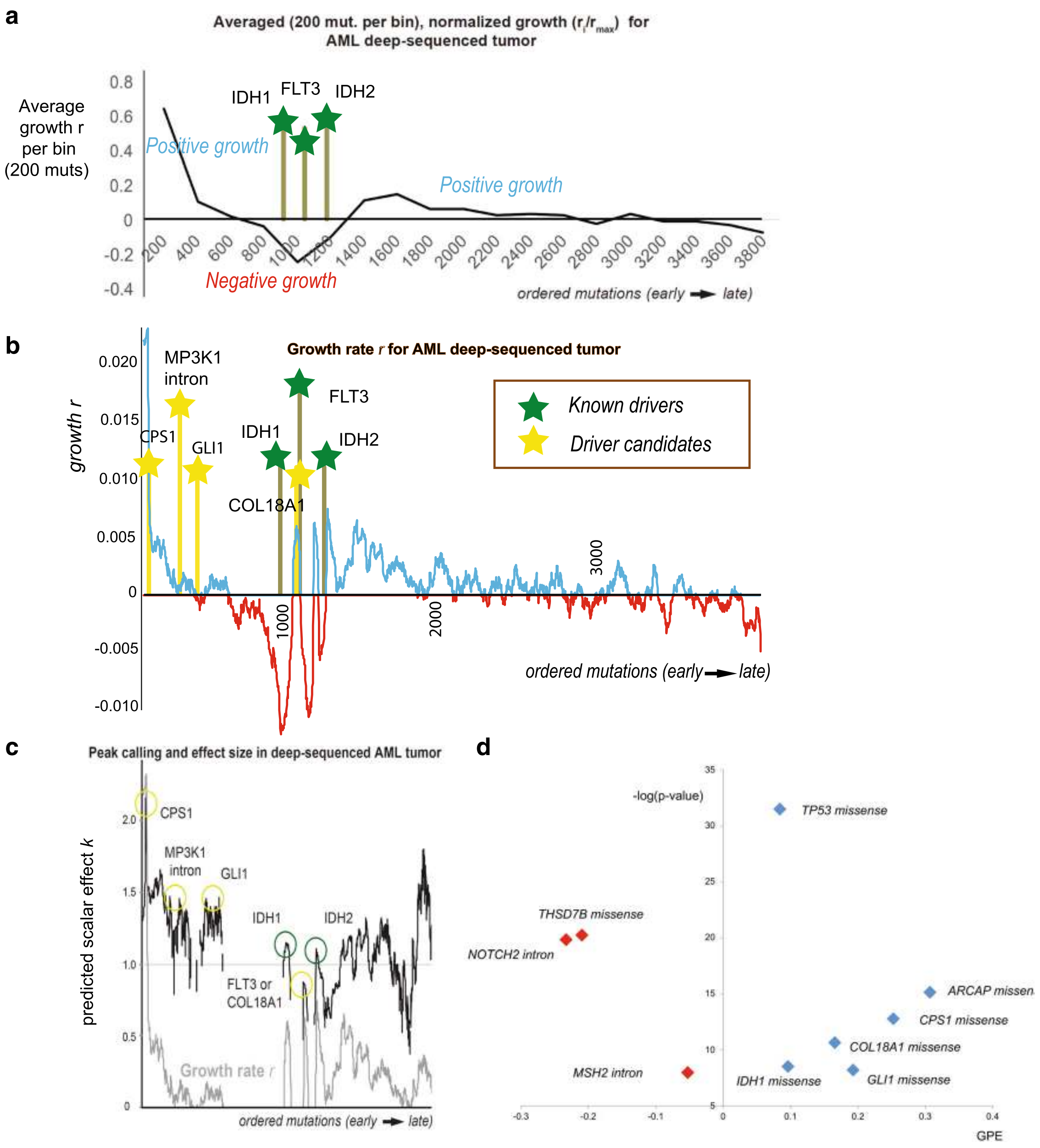


Fig. 5 Tumor progression on model AML liquid tumor. In **a**, we show the averaged growth progression for an AML ultra-deep-sequenced tumor. We ordered the sample's mutations from highest to lowest frequency and divided them into bins of 200 mutations. Three cancer mutations hit the tumor to establish a permanent growth (cancer mutations denoted by green bars). In **b**, we plot the mutational growth r_{i-1} for each mutation across tumor progression. The three cancer genes (*IDH1* missense, *FLT3* missense, and *IDH2* missense) aligned well with 3 of our top 5 growth peaks (two-tailed *t* test $p < 2.2e-16$). Candidate driver mutations -denoted by yellow bar—that we identified from our PCAWG database as being associated with positive growth (see also "(d)") aligned well with early—previously unjustified growth peaks. In **c**, we show each mutation's effect in tumor progression. Effect peaks corresponds to putative drivers. **d** By using our PCAWG database from our previous analysis, we tested which mutations from the deep-sequenced sample were associated with positive growth. The *x*-axis represents positive growth enrichment, while the *y*-axis shows the level of significance as the negative logarithm of a two-tailed *t* test *p* value ($-\log(p$ value$) > 5$. Overall, we found 6 mutation types that showed significant positive enrichment across 993 PCAWG tumors, including *TP53* missense (appeared during metastasis), *IDH1* missense, *COL18A1* missense, *CPS1* missense, *GLI1* missense, and *SRCAP* missense. Missense *TP53* and *SRCAP* mutations are not included in graph (b) as they were metastatic mutations. For association with positive growth we tested all missense mutations (e.g., *CPS1* missense), and every mutation in the sample from Vogelstein cancer genes (e.g., *NOTCH2* intron).

coverage, tumor ploidy, subclonality, and a low number of tumor samples per cancer or mutational element. Under a neutral model, our method would still detect some growth peaks or suggest the presence of weak drivers. These are false-positive predictions, possibly due to noise which results in various signal perturbations in the VAF spectrum, or potential genetic drift. Moreover, our model does not consider the potential effects from deleterious passenger mutations or sequencing errors on the VAF spectrum. However, we consider that -if not depleted- most deleterious mutations should have a small VAF in our sequenced sample. Similarly, we expect that sequencing errors tend to produce spurious mutations of extremely low VAF, which are ignored by our framework. Although some researchers are skeptical of the plausibility of VAF quantification[19,59], recent analyses have also confirmed that it can be achieved even at low sequencing coverage[16]. At the same time, as sequencing cost decreases exponentially, ultra-deep whole-genome sequencing for a larger number of samples will become trivially within reach. This is critical for the personalized assessment and parametrization of single samples.

Similar to previous Darwinian, bacterial, and viral evolution analyses, modeling the variations of cell populations allows us to associate these variations with specific events, even at a single sample level. Our work contributes to our understanding of cancer evolution by directly assessing tumor sample progression at the time of the driver event. This assessment can be very critical for therapeutic strategies and drug selection[32,33]. Our framework presents opportunities for personalized diagnosis via modeling the tumor's progression using deep-sequenced whole-genome data from one single individual.

Methods

Modeling the frequency of g-hitchhiking mutations. Let us assume that a simple population of cancer cells grows exponentially; for simplicity, we assign one mutation per cell division.

$$N(t) = \mathrm{e}^{rt}. \tag{4}$$

At sequencing time T, the frequency of a mutation occurring at time t_n would be equal to

$$f_\mathrm{n}(T, t_n) = \frac{\mathrm{e}^{r(T-t_\mathrm{n})}}{\mathrm{e}^{rT}} = \mathrm{e}^{-rt_\mathrm{n}}. \tag{5}$$

At time t_1, a mutation occurs that increases the growth rate r of the specific subpopulation by k, such that the new population is now expanding as

$$N_\mathrm{F} = \mathrm{e}^{krt}. \tag{6}$$

Thus, at total time $T = t_1 + t_2$, we expect a generational (g—) hitchhiking mutation that occurred at time $t_\mathrm{m} < t_1$ (see Fig. 1) to have a frequency equal to

$$f_\mathrm{g}(T, t_\mathrm{m}) = \frac{\mathrm{e}^{r(T-t_m)} + N_\mathrm{F} - \mathrm{e}^{rt_2}}{N_\mathrm{tot}}, \tag{7}$$

were N_tot is the total number of cells (or mutations) and N_F is the number of cells that contain the fitness mutation that occurred at t_1 and expanded for t_2.

$$\text{Thus, } N_\mathrm{F} = e^{krt_2}. \tag{8}$$

or

$$\mathrm{e}^{rt2} = \sqrt[k]{N_\mathrm{F}},$$

and Eq. (7) can be re-written as

$$f_\mathrm{g}(\mathrm{T}, \mathrm{t_m}) = \frac{\mathrm{e}^{\mathrm{r}(T-t_\mathrm{m})} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}}. \tag{9}$$

Moreover, if we assume that $t_\mathrm{m} \sim t_1$

then $f_\mathrm{g}(T, t_\mathrm{m} \sim t_1) = \frac{e^{r(t_1+t_2-t_\mathrm{m})} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} = \frac{\mathrm{e}^{rt_2} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} = \frac{N_\mathrm{F}}{N_\mathrm{tot}}$,or

$$\lim_{\mathrm{t_m} \sim \mathrm{t_1}} f_\mathrm{g}(T, t_\mathrm{m}) = \frac{N_\mathrm{F}}{N_\mathrm{tot}} = f_\mathrm{d}(T, t_1). \tag{10}$$

Eq. (9) can be rewritten as

$$\begin{aligned} f_\mathrm{g}(T, t_\mathrm{m}) &= \frac{\mathrm{e}^{r(t_1+t_2-t_\mathrm{m})} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} \Rightarrow \\ f_\mathrm{g}(T, t_\mathrm{m}) &= \frac{\mathrm{e}^{r(t1+t2)} \times \mathrm{e}^{-rt_\mathrm{m}} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} \Rightarrow \\ f_\mathrm{g}(T, t_\mathrm{m}) &= \frac{(N_\mathrm{tot} - N_\mathrm{F} + \mathrm{e}^{rt_2}) \times \mathrm{e}^{-rt_\mathrm{m}} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} \Rightarrow \\ f_\mathrm{g}(T, t_\mathrm{m}) &= \frac{\left(N_\mathrm{tot} - N_\mathrm{F} + \sqrt[k]{N_\mathrm{F}}\right) \times \mathrm{e}^{-rt_\mathrm{m}} + N_\mathrm{F} - \sqrt[k]{N_\mathrm{F}}}{N_\mathrm{tot}} \end{aligned}, \tag{11}$$

which is the frequency function f_g for the g-hitchhiking mutations in our sample.

Finally, according to (10) we get

$$f_\mathrm{g}(\mathrm{T}, \mathrm{t_m}) = \frac{\mathrm{e}^{-rt_\mathrm{m}} \times \left[N_\mathrm{tot} - f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot} + \sqrt[k]{f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot}}\right] + f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot} - \sqrt[k]{f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot}}}{N_\mathrm{tot}}, \tag{1}$$

We note that the frequency $f_\mathrm{g}(T, t_\mathrm{m})$ of g-hitchhiking mutations also follow the form of an exponential function.

$$f_\mathrm{g}(T, t_\mathrm{m}) = \mathrm{A} \times \mathrm{e}^{-rtm} + \mathrm{B}.$$

This allows the sampling and estimation of growth rate r from consecutive g-hitchhiking mutations $\mathrm{m_1}$, $\mathrm{m_2}$, and $\mathrm{m_3}$ that happened at corresponding times t_m1, t_m2, and t_m3 according to

$$r = \ln\left(\frac{f_\mathrm{g}(T, t_\mathrm{m1}) - f_\mathrm{g}(T, t_\mathrm{m2})}{f_\mathrm{g}(T, t_\mathrm{m2}) - f_\mathrm{g}(T, t_\mathrm{m3})}\right). \tag{2}$$

Independent calculation of growth *r* for g-hitchhikers. For three hitchhiking mutations that occured at times t, $t + n$, and $t + m$ ($n < m$), their respective frequencies are $f(t)$, $f(t + n)$, and $f(t + m)$

$$\text{Let } \Lambda = \frac{f_\mathrm{g}(T, t) - f_\mathrm{g}(T, t + n)}{f(T, t) - f_\mathrm{g}(T, t + m)} = \frac{(1 - e^{-rn})}{(1 - e^{-rm})}. \tag{12}$$

Thus,

$$\Lambda \times e^{-r \times m} - e^{-r \times n} - \Lambda + 1 = 0.$$

If we set $e^{-r} = x$
then

$$\Lambda \times x^m - x^n - \Lambda + 1 = 0.$$

By selecting $m = 2 \times n$

$$\Lambda \times \mathrm{x}^{2n} - x^n - \Lambda + 1 = 0.$$

Therefore

$$\begin{aligned} x^n &= \frac{1 \pm \sqrt{1-4*\Lambda(-\Lambda+1)}}{2^*\Lambda} = e^{-r \times n} \\ x &= e^{-r} = \sqrt[n]{\frac{1 \pm \sqrt{1-4*\Lambda(-\Lambda+1)}}{2*\Lambda}} \\ r &= -\log\left(\sqrt[\mathrm{n}]{\frac{1 \pm \sqrt{1-4*\Lambda(-\Lambda+1)}}{2*\Lambda}}\right) \end{aligned} . \tag{13}$$

Optimizing for t_g at any time point during progression. In (1), we associated g-hitchhiker frequency with population growth r as

$$f_\mathrm{g}(T, t_\mathrm{m}) = \frac{\mathrm{e}^{-rt_\mathrm{m}} \times \left[N_\mathrm{tot} - f_{\mathrm{d(T,t_1)}} \times N_\mathrm{tot} + \sqrt[k]{f_{\mathrm{d(T,t_1)}} \times N_\mathrm{tot}}\right] + f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot} - \sqrt[k]{f_{\mathrm{d}(T,t_1)} \times N_\mathrm{tot}}}{N_\mathrm{tot}}.$$

Furthermore, we included an extra parameter for "generational time" (t_g) allowing us to optimize for the number of generations until that point without knowledge of previous mutations.

Thus

$$f_\mathrm{g}\left(T, t_\mathrm{g}, t_\mathrm{i} - \mathrm{m}\right) = \frac{\mathrm{e}^{-r\left(t_\mathrm{g}+t_\mathrm{i}-m\right)} \times \left(N_\mathrm{tot} - f_{\mathrm{d}(T,t_i)} \times N_\mathrm{tot} + \sqrt[k_\mathrm{i}]{f_{\mathrm{d}(T,t_i)} \times N_\mathrm{tot}}\right) + f_{\mathrm{d}(T,t_\mathrm{i})} \times N_\mathrm{tot} - \sqrt[k_\mathrm{i}]{f_{\mathrm{d}(T,t_\mathrm{i})} \times N_\mathrm{tot}}}{N_\mathrm{tot}}, \tag{3}$$

where $f_\mathrm{d}(T, t_\mathrm{i})$ is the frequency of the putative driver i occurring at time t_i.

This allows us to reoptimiIe t_g at any time t_i during tumor growth independent of earlier calculations.

This allows us to

1. *exclude* pretumor somatic mutations or duplications in our calculations,
2. reoptimize at any time t during tumor growth independent of earlier calculations.

Behavior of $\hat{r}$ estimators on nongenerational mutations. We used the coalescent theory to analyze the behavior of the estimator

$$\hat{r} = \log\left(\frac{f_\mathrm{g}(T, t) - f_\mathrm{g}(T, t + n)}{f_\mathrm{g}(T, t + n) - f_\mathrm{g}(T, t + m)}\right), \tag{14}$$

when the assumption that the mutations are generational is not satisfied. Our results indicate that the growth indicator does not change qualitatively. We first analyzed the behavior in a constant-size population, and then in populations with increasing and decreasing exponential growth.

Behavior of $\hat{r}$ estimators for constant population sizes. We consider a population of constant size N with mutation rate μ. Given the population observed at a fixed time point t_0, we can consider the coalescent tree of all cells at $t = 0$ reaching back to their most recent common ancestor at $t = T$ (where time is indexed in reverse direction from t_0). Writing T_n for the length of time over which n lineages are present (i.e., the time between the splits of lineage n and $n + 1$, hence $T = \sum_{n=1}^{N} T_\mathrm{n}$, where T_N is truncated at time 0), using Kingman's coalescent[60] it can be shown that

$$T_n = \frac{2}{n(n-1)}. \tag{15}$$

Assuming that the birth-and-death rates remain constant, the number of mutations M_{n} acquired during T_{n} can be expressed up to a constant of proportionality

$$M_n \propto \mu n T_{\mathrm{n}} = \frac{2\mu}{n-1}. \tag{16}$$

We can approximate the variation in density of the VAF spectrum by assuming that all mutations falling in T_{n} take their expected frequency $f = 1/n$, and calculating the density δ_{n} within windows $[1/n1/(n-1))$, whose lengths are $L_{\mathrm{n}} = 1/(n-1) - 1/n = 1/(n(n-1))$

$$\delta_n = \frac{M_n}{L_n} \propto 2\mu n. \tag{17}$$

As δ_{n} increases with n, and hence with time, $\hat{r}$ as estimated using Eq. (14) is predicted to take positive values in a constant-size population.

Populations with varying size. To predict the behavior of $\hat{r}$ in populations of varying size, we formulated the coalescent as a Markov process. We let $t = 0$ represent the time of observation and indexing time in reverse as above, and X_{t} for a random variable represent the number of lineages in the coalescent tree at time t, and N_{t} for the population size at time t. To define a Markov process over X_{t} from $t = 0\ldots T$, we fixed the initial distribution to $p(x_0 = i) = [i = N]$, where [.] is the Iverson bracket. We also defined a transition matrix τ such that $\tau(i,j)$ represents the conditional probability $p(x_{\mathrm{t+1}} = j|x_{\mathrm{t}} = \mathrm{i})$; that is, the probability that there are j coalescent lineages at time $t + 1$ given there are i lineages at time t. Note that the number of coalescent lineages present will be less than or equal to the size of the population at a given time.

To calculate $\tau(i, j)$, we evaluated the number of maps that take i lineages to j lineages given the final population size of $N_{\mathrm{t+1}}$. As there are j final lineages, the image of the map must have size j, meaning that it must be one of the $\begin{pmatrix} N_{\mathrm{t+1}} \\ j \end{pmatrix}$ subsets of the population at $t + 1$. For each of these subsets, the original i lineages can be partitioned into those taking distinct values at $t + 1$, so that there are I possible partitions of i lineages into j nonempty subsets, where $\left\{ \begin{matrix} i \\ j \end{matrix} \right\} = \frac{1}{j!}\sum_{k=0}^{j}(-1)^{j-k}\begin{pmatrix} j \\ k \end{pmatrix}k^n$ is a Stirling number of the second kind. Further, there are $(j!)$ permutations of the image set to which each of these partitions caI be mapped. Given $N^{\mathrm{i}}_{\mathrm{t+1}}$ maps in total from the populations from t to $t + 1$ when restricted to the ancestors in the coalescent tree

$$\boldsymbol{\tau}(i,j) = \frac{\begin{pmatrix} N_{t+1} \\ j \end{pmatrix}\left\{ \begin{matrix} i \\ j \end{matrix} \right\}(j!)}{N^{i}_{t+1}} = \begin{pmatrix} N_{t+1} \\ j \end{pmatrix}\frac{\sum_{k=0}^{j}(-1)^{j-k}\begin{pmatrix} j \\ k \end{pmatrix}k^n}{N^{i}_{tI+1}}I, \tag{18}$$

assuming $j \le i$, and $\tau(i, j) = 0$ otherwise.

To investigate the behavior of $\hat{r}$, we used the Markov chain above to calculate $p_{\mathrm{t+1}} = p_{\mathrm{t}}\tau$ for a fixed number of time-steps, where $p_{\mathrm{t}} = [p(x_{\mathrm{t}} = 1), p(x_{\mathrm{t}} = 2), \ldots, p(x_{\mathrm{t}} = N)]$. We then calculated a function $g(t)$ representing the expected number of lineages present at time t, $g(t) = \sum_{\mathrm{n}} np(x_{\mathrm{t}} = n)$. We calculated T_{n}, which estimates the length of time over which there are exactly n lineages as above as

$$T_n = \min(\{t|g(t) \le n\}) - \min(\{t|g(t) \le n-1\}), \tag{19}$$

from which the total number of mutations M_{n} acquired during T_{n} can be calculated using Eq. (16), and the variation in density of the VAF spectrum over windows corresponding to the intervals T_{n} can be calculated using Eq. (17).

We calculated δ_{n} as a function of n in a number of populations, using the population model $N^{\mathrm{t+1}} = \alpha N^{\mathrm{t}}$, where we let $\alpha = [1,1.1,1.2,\ldots,2]$, corresponding to a decreasing population (as time is indexed in reverse), and $\alpha = [1,0.9,0.8,\ldots,0.5]$, corresponding to an increasing population. We used 200 time-steps for all calculations with $\mu = 0.01$. We started all decreasing populations at $N_0 = 10$ and fixed a maximum population size of $N_{\max} = N_0\alpha^{10}$, while fixing a minimum size for all increasing populations at $N_{\min} = 10$, and starting at $N_0 = N_{\min}\alpha^{-10}$. For all t after the population reaches its maximum/minimum size, we set $N^{\mathrm{t+1}} = N^{\mathrm{t}}$. Supplementary Fig. 2 shows the output for populations of decreasing and increasing sizes, respectively. As predicted by the earlier analysis, the calculations show that δ_{n} is an increasing function for constant population size ($\alpha = 1$), corresponding to a positive value of $\hat{r}$, and is approximately linear. Likewise, δ_{n} is increasing for all populations of increasing size (Supplementary Fig. 2b); hence, $\hat{r}$ is predicted to be positive for all such populations, with a magnitude increasing with α, as the rate of increase of δ_{n} increases for larger α. For decreasing populations (Supplementary Fig. 2a), δ_{n} is only strictly decreasing for $\alpha > 1.4$, corresponding to $r = -\log(\alpha) \approx -0.34$ in generational units, suggesting that a negative r that is at least this magnitude will result in a negative estimate for $\hat{r}$ using Eq. (14).

Reconsidering the assumption of one mutation per division. In our model, we have assumed for reasons of convenience and simplicity that one new mutation arises per cell division. However, this assumption is not required to implement our model. To derive the estimator for r in Eq. (2), all that is required is that the intervals $t_{\mathrm{m2}} - t_{\mathrm{m1}}$ and $t_{\mathrm{m3}} - t_{\mathrm{m2}}$ are equal in expectation. For a mutation rate 0.5 \μ = 1 (where \μ is the total number of mutations expected per cell division), this interval is one generation, but for \μ < 2 the expected interval is 2/\μ.

Model optimization and initial parameters. To optimize our model, we used custom perl scripts and the R package "Nonlinear Least Squares" (NLS)[61,62] with sliding windows of $m = 150$ g-hitchhikers. We optimized for [["mod < −nls($\overrightarrow{f}$ ~ exp (−r * (t_{g} + $\overrightarrow{mut\_order}$))*(1 − α) + α, start = list(α = 0.01, t_{g} = 1), control = nls.control(maxiter = 10,000,000, tol = 1e−04, minFactor = 0.000002, printEval = TRUE, warnOnly = TRUE))"]], where f is the frequency vector for the g-hitchhickers, t_{g} corresponds to generational time, and α is a composite parameter associated with the driver's prevalence and its respective effect according to equation

$$f_{\mathrm{g}}\left(T, t_{\mathrm{g}}, t_{\mathrm{i}} - m\right) = \frac{\mathrm{e}^{-r\left(t_{\mathrm{g}}+t_{\mathrm{i}}-m\right)} \times \left(N_{\mathrm{tot}} - f_{\mathrm{d}(T,t_i)} \times N_{\mathrm{tot}} + \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}}}\right) + f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}} - \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}}}}{N_{\mathrm{tot}}},$$

where $f_{\mathrm{d}}(T, t_{\mathrm{i}})$ is the frequency of the putative driver i occurring at time t_{i}.

Growth r can be estimated using equation

$$r = \ln\left(\frac{f_{\mathrm{g}}(T, t_{\mathrm{m1}}) - f_{\mathrm{g}}(T, t_{\mathrm{m2}})}{f_{\mathrm{g}}(T, t_{\mathrm{m2}}) - f_{\mathrm{g}}(T, t_{\mathrm{m3}})}\right),$$

or alternatively through NLS optimization.

When the population shows an exponential growth and we are only interested in the first driver or the subclonal effect, we can omit the generational time (t_{g}) estimation to reduce unnecessary optimizing errors. In this case, we simplify the R command

[["mod < - nls($\overrightarrow{f}$ ~ exp(−r * $\overrightarrow{mut\_order}$)*(1 − α) + α, start = list(α = 0.01), control = nls.control(maxiter = 10000000, tol = 1e-04, minFactor = 0.000002, printEval = TRUE, warnOnly = TRUE))"]] or [["mod < - nls($\overrightarrow{f}$ ~ exp(−r * $\overrightarrow{mut\_order}$) × α + β, start = list(α = 0.01, β = 0.1), control = nls.control(maxiter = 10000000, tol = 1e-04, minFactor = 0.000002, printEval = TRUE, warnOnly = TRUE))"]].

Finding the best sliding window size m. For our PCAWG analysis, we used a sliding window of $m = 100$ and 150 g-hitchhikers. Our presented results are based on $m = 150$. Overall, larger windows provided a more stable uniformal analysis across all 993 tumors, allowing the NLS algorithm to converge easier across all samples, by considering a larger range of mutation frequencies in tumors with lower coverage. However, especially in deep-sequenced tumors, the size of a sliding window should be individually optimized based on population assumptions. For our simulation analyses, we selected the size of a sliding window that minimizes the absolute median ranking distance $|\widetilde{|D|}$ between true and predicted drivers by maximizing the p value when compared to a distance distribution of 100 random mutations. For our independent set of nonneutral simulations based on Williams et al. software, we were able to calculate an optimal window size by tuning our algorithm based on 464 nonneutral simulations. The optimal window size that provided a median effect of 1 for the neutral simulations was 150 hitchhikers, which is also what we used for the deep-sequenced AML tumor. Smaller window sizes provided a higher median effect for both neutral and nonneutral simulations, without burdening our method's detectability.

Scalar *k* and selection coefficient *s*. In real tumors, cells bearing a subclonal driver mutation can form a distinguishable subclone within a tumor of millions, billions or trillions of cells, as a result of small growth advantage of these cells compounded over hundreds to thousands of generations. The k values used in our simulations should therefore be scaled when predicting the corresponding k^* values in a real population on which our estimators would exhibit similar behavior (due to similar amounts of variance/genetic drift). Felsenstein[60] describes scaling rules for simulations: to use a smaller population to simulate a larger one, the quantity 4 × $N \times s$, where N is the population size and s the selection coefficient must remain the same. We consider a range of population sizes (10^6–10^{10}) as being realistic (Williams et al.[16] use an estimate of 10^{10} cells; we note however that spatial effects may result in a lower effective population size in many tumors). Using the scaling $k* = 1 + \frac{N_s(k-1)}{N_{\mathrm{r}}}$, where N_{s} and N_{r} are the simulation and realistic population sizes respectively, the range of k we considered in our simulations from 1.1 to 4, corresponds under this scaling to k^* \in [1.001, 1.03], which are noticeably smaller than 1.1 ($s^* = 0.1$ corresponding to a very strong driver effect). We consider these values to be upper-bounds, as the true effective population size is likely to be substantially larger than one million cells for most cancers. As an alternative to scaling the values of k as discussed, we also consider the effects of directly substituting realistic size estimates (10^6 to 10^{10}) into the variable N_{tot} in equation $f_{\mathrm{g}}\left(T, t_{\mathrm{g}}, t_{\mathrm{i}} - m\right) = \frac{\mathrm{e}^{-r\left(t_{\mathrm{g}}+t_{\mathrm{i}}-m\right)} \times \left(N_{\mathrm{tot}} - f_{\mathrm{d}(T,t_i)} \times N_{\mathrm{tot}} + \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}}}\right) + f_{\mathrm{d}(T,t_{\mathrm{i}})} \times N_{\mathrm{tot}} - \sqrt[k_{\mathrm{i}}]{f_{\mathrm{d}(T,t_{\mathrm{i}})} \times \mathrm{N}_{\mathrm{tot}}}}{\mathrm{N}_{\mathrm{tot}}}$. As shown in Supplementary Fig. 1, this leads to an improvement in the accuracy with which we detect simulated drivers (in terms of the distance from the simulated driver). Our simulations thus imply that our algorithm can detect drivers with weak effects accurately in tumors of realistic sizes.

Simulating tumors of lower coverage depth in sequencing. To simulate sequencing coverage depth of 100×, 300×, 500×, and 700×, we first created a cell population from 1000× coverage based on mutational frequencies. In this sense, a mutation with a frequency of 0.475 would be associated with 475 out of 1000 individuals that contained the specific mutation (noted as "1") and 525 individuals that did not (noted as "0"). Consequently, we sampled with replacement sub-populations with sizes 100, 300, 500, and 700 cells. Then we recalculated the lower coverage frequencies as the sum of "1"s divided by the corresponding coverage.

True positives, false positives, sensitivity, and PPV analysis. To test our model's performance across simulations with various driver effects and depth coverage we aimed to determine the number of true positive (TP), false positive (FP), true negatives (TN) and false negative (FN) predictions. Based on our previous calculations, we estimated the absolute median distance ($|\widetilde{|D|}$) between the true and the predicted driver's position close to 11 mutations and the standard error (SE) about 2.5 mutations. For every simulation, a driver prediction was considered as TP, if $|\widetilde{|D|}$ between our predicted driver and the true driver was less than $|\widetilde{|D|} + 2\mathrm{SE} = 16$ mutations. If the predicted driver peak was at distance longer than 15 mutations, the corresponding predicted driver was considered as FP. If a simulation provided zero TPs, we then considered the lack of TP as FN. Similarly, if a simulation resulted in zero FPs, we then considered the lack of FP as a TN. We should note that the cut off of 15 mutations is fairly strict for the detection of TPs, especially for a real size tumor/samples, but helpful to systematically evaluate our model across different simulations.

Ranking distance *D* (between true driver and effect peak). For each simulation, we calculate the ranking distance D (as a number of ordered mutations) between the ranked position of simulated driver and the predicted growth or effect peak (predicted driver). We further calculate the median $\widetilde{D}$ and absolute median distance $\widetilde{|D|}$ for different k effects, across simulated sequencing coverage, population size, etc. To assess statistical significance for our method's ability to detect drivers, for each simulation, we also draw a random driver prediction. Then, for each sample we calculate the distance between our random prediction and the simulated driver as previously described. Finally, we compare the median, absolute median, standard deviation, and SE against the random distance distribution for all simulations. p Values are obtained from a two-tailed t-test. In samples where the true driver is unknown (e.g., PCAWG linear tumors), the ranking distance D from a growth peak corresponds to the ranked number of order mutations for one specific mutation from the closest peak (e.g., PCAWG drivers).

Positive growth enrichment. A type of mutation (e.g., *TP53* missense) is assessed if it occurs significantly more often than random during periods of positive growth r. For every type of mutation that we tested, we picked an equal number of random mutations from the same individual samples. By repeating this process 100 times we set our mean expectation for randomly associating mutations with positive growth. Then we assess if the specific type of mutation is enriched during periods of positive growth compared to random. For example, for $n = 48$ "*TP53* missense" mutations in our sample with positive growth, we found a comparative average of $\overline{x_{\mathrm{pr}}} = 40.38$ of random mutations with a SE of mean SEM = 0.42. Significance was then assigned based on z-score. As PGE we report the value of $\mathrm{PGE} = \frac{(x_{\mathrm{p}} - \overline{x_{\mathrm{pr}}})}{\mathrm{total\#(of\ e.g.TP53\ missense)}}$, where in this example, x_p is the number of *TP53* missense mutations found during positive growth and $\overline{x_{\mathrm{pr}}}$ is the average number of random mutations with positive growth r, as sampled 100 times with replacement.

Significance for deep-seq AML tumor drivers. To test the level of significance for our growth peak prediction in the deep-sequenced tumor, we selected our top five highest growth peaks and estimated the distance D between the three known cancer genes and the closest growth peak. Then, for 1000 replicates we sampled random mutations with replacement to create a random distribution of distances between a random mutation and its closest peak. For a more conservative approach, we increased the number of highest peaks to ten and reduced the random mutation sample to the first 2000 mutations without losing significance. A two-tailed t test was performed to establish the level of significance.

Enrichment across effect bins. To estimate enrichment across different effect ranges for specific mutation types (e.g., *TP53* missense mutations) we created effect bins of $k = [0.9–1.1, 1.1–1.3, \ldots\ 2.1–2.3, 2.3–2.5, 2.5–2.7]$. Across the PCAWG samples, for each mutation, we also picked one random mutation from the same sample. Then, we bootstrapped this process for 100 replicates. Finally, for every bin we tested whether the specific mutation appeared to be enriched compared to random.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

PCAWG protected datasets are controlled access that is subject to data usage agreement. PCAWG datasets are available upon request and authorization from the ICGC Data Access Compliance Office and dbGaP Authorized Access program for US-based projects. For data repositories and data request see https://docs.icgc.org/pcawg/data/. Pseudo-VCF files are provided at https://doi.org/10.6084/m9.figshare.9722651.v1. These files contain real VAF distributions, mutation type information including gene names, but all genomic coordinates and variance information have been masked and randomly modified. The source data underlying Figs. 2–5 and Supplementary Figs are provided as a Source Data file.

Code availability

We are providing a perl script that analyzes a pseudo-VCF derived file format for growth and effect calculation. It should be noted that our model does not correct for purity and ploidy inconsistencies, but instead utilizes already derived mutational frequencies. Our code is publicly available, together with test play data and a readme file at https://github.com/gersteinlab/Evotum101.

Received: 10 October 2018; Accepted: 26 November 2019;
Published online: 05 February 2020

References

1. Heng, H. H. Q. et al. Stochastic cancer progression driven by non-clonal chromosome aberrations. *J. Cell. Physiol.* **208**, 461–472 (2006).
2. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat. Rev. Cancer* **11**, 450–457 (2011).
3. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
4. Navin, N. et al. Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20**, 68–80 (2010).
5. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
6. Yngvadottir, B., Macarthur, D. G., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. *Genome Biol.* **10**, 237 (2009).
7. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K. & Gerstein, M. B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res.* **39**, 7058–7076 (2011).
8. Bernstein, B. E. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
9. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
10. Haussler, D. et al. *A Million Cancer Genome Warehouse* (2012).
11. Campbell, P. J. et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.0801523105 (2008).
12. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* https://doi.org/10.1016/j.cels.2015.08.015 (2015).
13. Krimmel, J. D. et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.1601311113 (2016).
14. Subramanian, A., Shackney, S. & Schwartz, R. Inference of tumor phylogenies from genomic assays on heterogeneous samples. *J. Biomed. Biotechnol.* **2012**, 797812 https://doi.org/10.1155/2012/797812 (2012).
15. Sottoriva, A. et al. A big bang model of human colorectal tumor growth. *Nat. Genet.* https://doi.org/10.1038/ng.3214 (2015).
16. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0128-6 (2018).
17. Turajlic, S. et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell.* https://doi.org/10.1016/j.cell.2018.03.043 (2018).
18. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. *Cell.* https://doi.org/10.1016/j.cell.2018.02.020 (2018).
19. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Reply: uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.* https://doi.org/10.1038/ng.3877 (2017).
20. Tarabichi, M. et al. Neutral tumor evolution? *Nat. Genet.* https://doi.org/10.1038/s41588-018-0258-x (2018).
21. Heide, T. et al. Reply to 'Neutral tumor evolution?'. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0256-z (2018).
22. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.1616440113 (2016).

23. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer*. https://doi.org/10.1038/nrc3999 (2015).
24. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* https://doi.org/10.1038/ng.3987 (2017).
25. Parmigiani, G. et al. Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics*. https://doi.org/10.1016/j.ygeno.2008.07.005 (2009).
26. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. https://doi.org/10.1038/nature12213 (2013).
27. Dees, N. D. et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* https://doi.org/10.1101/gr.134635.111 (2012).
28. Vogelstein, B. et al. Cancer genome landscapes. *Science*. https://doi.org/10.1126/science.1235122 (2013).
29. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gks743 (2012).
30. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt395 (2013).
31. Davoli, T. et al. XCumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*. https://doi.org/10.1016/j.cell.2013.10.011 (2013).
32. Reiter, J. G. et al. Minimal functional driver gene heterogeneity among untreated metastases. *Science*. https://doi.org/10.1126/science.aat7171 (2018).
33. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* https://doi.org/10.1186/s13073-018-0531-8 (2018).
34. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1421839112 (2015).
35. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer*. https://doi.org/10.1038/35101031 (2001).
36. Diederichs, S. et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol. Med.* https://doi.org/10.15252/emmm.201506055 (2016).
37. Kumar, S. et al. Passenger mutations in 2500 cancer genomes: overall molecular functional impact and consequences. *bioRxiv*. https://doi.org/10.1101/280446 (2018).
38. Soussi, T., Taschner, P. E. M. & Samuels, Y. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. *Hum. Mutat.* https://doi.org/10.1002/humu.23163 (2017).
39. Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L. & Samuels, Y. The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment Cell Melanoma Res.* https://doi.org/10.1111/pcmr.12413 (2015).
40. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*. https://doi.org/10.1016/j.cell.2014.01.051 (2014).
41. Chen, Y. T. et al. Tumor-associated intronic editing of HNRPLL generates a novel splicing variant linked to cell proliferation. *J. Biol. Chem.* https://doi.org/10.1074/jbc.RA117.001197 (2018).
42. Jung, H. et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* https://doi.org/10.1038/ng.3414 (2015).
43. Hurst, L. D. & Batada, N. N. Depletion of somatic mutations in splicing-associated sequences in cancer genomes. *Genome Biol.* https://doi.org/10.1186/s13059-017-1337-5 (2017).
44. Gartner, J. J. et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.1304227110 (2013).
45. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
46. Castellanos-Moreno, A., Castellanos-Jaramillo, A., Corella-Madueño, A., Gutiérrez-López, S. & Rosas-Burgos, R. Stochastic model for computer simulation of the number of cancer cells and lymphocytes in homogeneous sections of cancer tumors. Preprint at https://arxiv.org/abs/1410.3768 (2014).
47. Turner, C., Stinchcombe, A. R., Kohandel, M., Singh, S. & Sivaloganathan, S. Characterization of brain cancer stem cells: a mathematical approach. *Cell Prolif.* https://doi.org/10.1111/j.1365-2184.2009.00619.x (2009).
48. Baar, M. et al. A stochastic model for immunotherapy of cancer. *Sci. Rep.* https://doi.org/10.1038/srep24169 (2016).
49. Figueredo, G. P., Siebers, P. O., Owen, M. R., Reps, J. & Aickelin, U. Comparing stochastic differential equations and agent-based modelling and simulation for early-stage cancer. *PLoS One*. https://doi.org/10.1371/journal.pone.0095150 (2014).
50. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* https://doi.org/10.1016/0021-9991(76)90041-3 (1976).
51. Székely, T. & Burrage, K. Stochastic simulation in systems biology. *Comput. Struct. Biotechnol. J.* https://doi.org/10.1016/j.csbj.2014.10.003 (2014).
52. Ryser, M. D., Lee, W. T., Ready, N. E., Leder, K. Z. & Foo, J. Quantifying the dynamics of field cancerization in tobacco-related head and neck cancer: a multiscale modeling approach. *Cancer Res.* **76**, 7078–7088 (2016).
53. Yates, C. A., Ford, M. J. & Mort, R. L. A multi-stage representation of cell proliferation as a Markov process. *Bull. Math. Biol.* https://doi.org/10.1007/s11538-017-0356-4 (2017).
54. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* https://doi.org/10.1038/ng.3489 (2016).
55. Celiktas, M. et al. Role of CPS1 in cell growth, metabolism, and prognosis in LKB1-inactivated lung adenocarcinoma. *J. Natl Cancer Inst.* https://doi.org/10.1093/jnci/djw231 (2017).
56. Kim, J. et al. CPS1 maintains pyrimidine pools and DNA synthesis in KRAS/LKB1-mutant lung cancer cells. *Nature*. https://doi.org/10.1038/nature22359 (2017).
57. Lo, H.-W., Zhu, H., Cao, X., Aldrich, A. & Ali-Osman, F. A novel splice variant of GLI1 that promotes glioblastoma cell migration and invasion. *Cancer Res.* https://doi.org/10.1158/0008-5472.CAN-09-0886 (2009).
58. Mastrangelo, E. & Milani, M. Role and inhibition of GLI1 protein in cancer. *Lung Cancer Targets Ther.* **9**, 35–43 (2018).
59. Noorbakhsh, J. & Chuang, J. H. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.* **49**, 1288 (2017).
60. Joseph Felsenstein, J. F. *Inferring Phylogenies*. (Sinauer Associates, Inc., 2004).
61. Bates, D. M. & Watts, D. G. *Nonlinear Regression Analysis and Its Applications*. (Wiley: New York, 1988) 10.1002/9780470316757.
62. Fox, J. & Weisberg, S. Nonlinear regression and nonlinear least squares in R. *R Compan. Appl. Regres*. 10.1016/j.infsof.2008.09.00 5 (2011).

Acknowledgements

We thank the PCAWG consortium for the state-of-the-art preliminary analysis and management of the data. We thank the members of the Gerstein lab Cancer Genomics group (S.K., S.L., and P.D.M.) for helpful discussion on the method development and analysis of the data. We acknowledge support from the NIH Grant (R01 HG 008126). This work was partially supported by NIH/NIGMS T32 GM007205.

Author contributions

L.S. conceived of the project, designed, performed, and analyzed the experiments. W.M. designed and developed the simulations. J.W. performed the coalescent analysis and designed the simulations. L.S. drafted the paper. L.S. and M.G. wrote the paper. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-14407-9.

Correspondence and requests for materials should be addressed to M.G.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.