Neutral tumor evolution?

Maxime Tarabichi[1], Iñigo Martincorena[2], Moritz Gerstung[3], Florian Markowetz[4], Paul T. Spellman[5], Quaid D. Morris[6], Ole Christian Lingjærde[7], David C. Wedge[8], Peter Van Loo[1,9,*], on behalf of the PCAWG Evolution and Heterogeneity Working Group[10]

[1]The Francis Crick Institute, London, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [3]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge, United Kingdom; [4]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; [5]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA; [6]University of Toronto, Toronto, Canada; [7]Department of Informatics and Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway; [8]Big Data Institute, University of Oxford, Oxford, United Kingdom; [9]Department of Human Genetics, University of Leuven, Leuven, Belgium.

[10]A list of members of the PCAWG Evolution and Heterogeneity Working Group can be found at the end of the manuscript.

[*]To whom correspondence may be addressed: The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, United Kingdom. Tel: +44 (0) 20 3796 1719, e-mail: Peter.VanLoo@crick.ac.uk.

Tumor growth is an evolutionary process governed by somatic mutation, clonal selection and random genetic drift[1]. Tumor subclones are subpopulations of tumor cells with a common set of mutations resulting from the expansion of a single cell during tumor development, and have been observed in a significant fraction of cancers and across multiple cancer types[2]. Akin to ongoing discussions in the field of speciation genetics, the relative importance of selection and genetic drift in the emergence of these tumor subpopulations remains unknown. According to a model proposed by Nowell[3], tumors evolve through a series of selective sweeps, whereby one cell acquires a selective advantage and its lineage becomes predominant. According to this traditional model, tumor subclones reflect ongoing selective sweeps. While the vast majority of mutations found in a tumor genome are passengers, a much smaller set of driver mutations is thought to provide a handle for natural selection concomitant with clonal expansion[1].

Williams *et al.*[4] recently claimed that under neutral evolution and given a simple model of tumor growth there would be a linear relationship between the number of passenger mutations $M(f)$ present in a fraction f of cells and the reciprocal of that fraction: $M(f) \propto \frac{1}{f}$. The authors argued that this relationship provides a convenient and intuitive null model, in which deviation indicates the presence of selection. In particular they note that the values required to assess it are routinely measured in DNA-sequencing data: mutations and their corresponding variant allele fractions (VAF), from which f can be derived. In real cancer data from The Cancer Genome Atlas (TCGA), Williams *et al.* reported no detectable deviation from the proposed linear relationship in about one third of the cases and concluded that these tumors are neutrally evolving. While providing an interesting approach to infer selection in human cancers, the analysis by Williams *et al.* is unfortunately limited by four major simplifying assumptions that render their conclusions questionable.

First, inferring f of variants from their VAF requires accurate estimates of local copy number, overall tumor purity and ploidy. Williams *et al.* tried to account for some of these factors by restricting their analyses to variants with VAF between 0.12 and 0.24 and located in copy-neutral regions of the genome. However, tumors with whole genome duplications, (i.e. 37% of tumors in the analyzed dataset[5]), have a clonal peak of mutations at or below VAF=0.25, which would lead to artificial deviation from the linear fit within that VAF window. The approach of Williams *et al.* would therefore likely produce false neutral calls in a significant fraction of tumors.

Second, the interpretation of the analyses is inconsistent with the use of neutrality as a null model. Failure to reject the null hypothesis is not the same as proving it is true. To infer neutrality one would need to demonstrate, in addition, either that a linear fit is sufficient to infer neutrality or the corollary, that all models of non-neutral tumor growth yield non-linear relationships.

Using the method described by Williams *et al.*, we fail to reject neutrality in simulated tumors in which we explicitly model subclonal growth with a selective advantage, i.e. increasing the division rate λ or the mutation rate μ of the subclone (**Supplementary Methods**). In fact, non-neutrality is detected only within a narrow range of λ and μ values tested that would lead to detectable subclones (true rejection of neutrality in ~11% of simulations; **Fig. 1a**). We conclude that a linear fit is not sufficient to call neutrality and that misuse of this model is likely to result in substantial over-calling of neutrality.

Third, the deterministic model of tumor growth described by Williams *et al.* relies on strong biological assumptions, among which are synchronous cell divisions, constant cell death and constant mutation and division rates. Stochastic models of tumor growth are biologically more realistic, as they allow for asynchronous divisions and probabilistic mutation acquisition, cell death and division rates. Using branching processes to simulate neutral and non-neutral growth[6] (**Supplementary Methods**), we show that $M(f) \propto \frac{1}{f}$ is neither a necessary nor a sufficient property of neutrally evolving tumors. Although it can be shown that the expected cumulative number of mutations – i.e. the average over many independent samples – $\overline{M}(f) \propto \frac{1}{f}$,[6] due to the biological noise modeled in branching processes, a typical realization of the neutral process in a single sample deviates substantially from the expected linear fit. As a result, discrimination of neutral and non-neutral simulated tumors using a linear fit is almost arbitrary, with 53.5% false positive neutral calls in non-neutral tumors (**Fig. 1b**) and an area under the ROC curve of 0.42 for the classification of 1,919 neutral and 1,919 non-neutral tumors (**Fig. 1c**).

Fourth, we reason that in tumors called neutral, no selection should be detected. To evaluate this, we use an orthogonal method to identify selection, based on the observed variants themselves rather than on their allele frequencies. dN/dS analysis derives the fraction of mutated non-synonymous positions to the fraction of mutated synonymous positions in the coding regions. It is widely used to detect the presence of negative or positive selection of non-synonymous variants in coding regions[7]. We applied a dN/dS model optimized for the

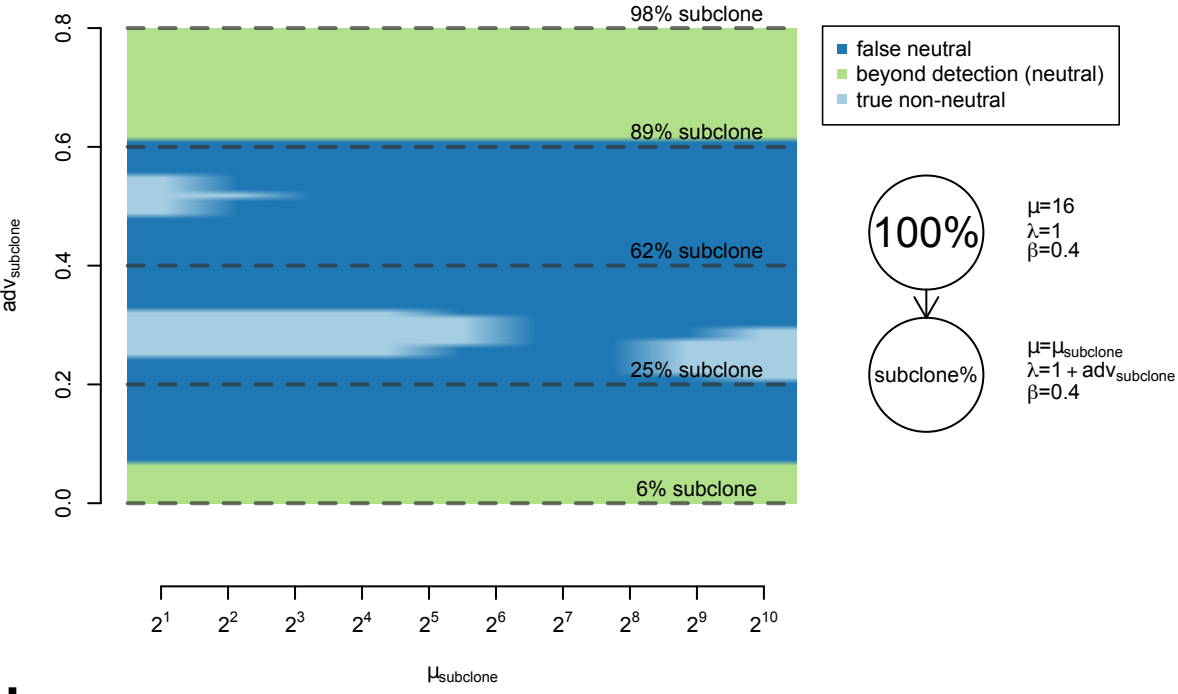
detection of selection in somatic cancer variants[7] to the TCGA exome data (**Supplementary Methods**). The analysis was performed separately using variants called as clonal or subclonal[8] (**Supplementary Methods**), in tumors called neutral and non-neutral based on the rationale outlined by Williams and collaborators[4]. dN/dS ratio analysis revealed significant positive selection in subclonal mutations of tumors classified as neutral (**Fig. 1d**), reinforcing the conclusion that the approach of Williams *et al.* is under-equipped to detect the presence or absence of selection.

It is of clinical importance to identify and better understand the drivers of the potentially more aggressive (sub)clones expanding under selective biological or therapeutic pressure, as these are good candidates for predicting resistance and exploring combination therapy. Williams *et al.* claimed that about one third of tumors are neutrally evolving. However, we find that their approach often leads to identification of individual tumors as neutral when they are non-neutral and non-neutral when they are neutral. Therefore quantification of selection during the evolution of single tumors using allele frequencies remains an open challenge.

a

deterministic neutral model $\Rightarrow M(f) \propto \frac{1}{f}$

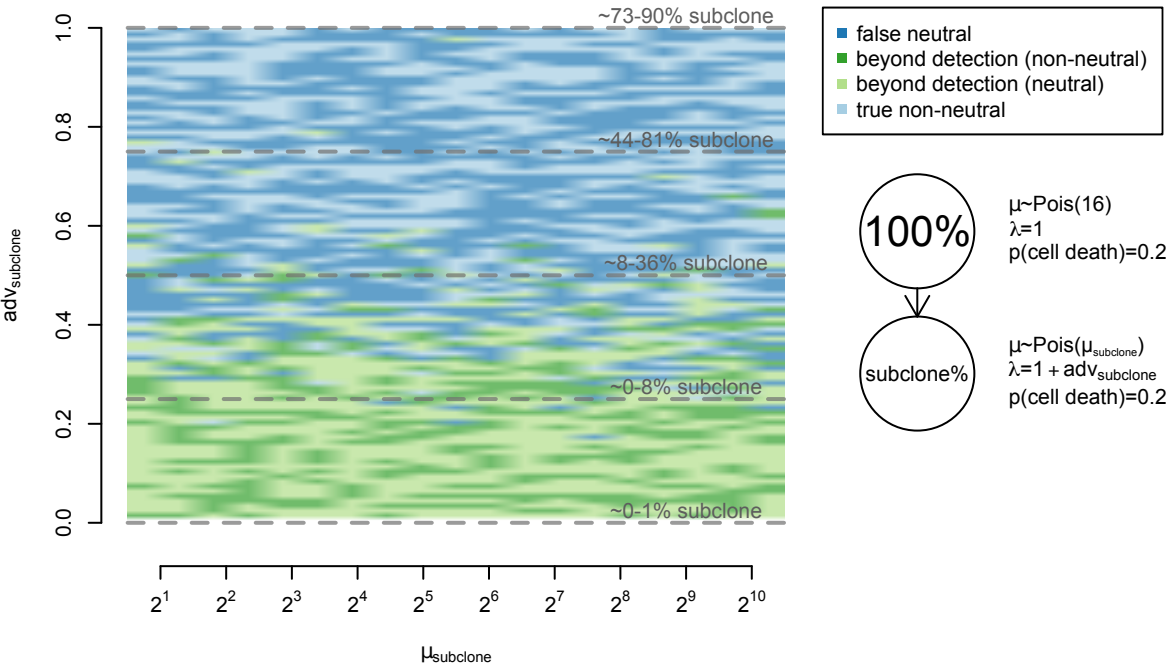
deterministic neutral model $\not\Leftarrow M(f) \propto \frac{1}{f}$



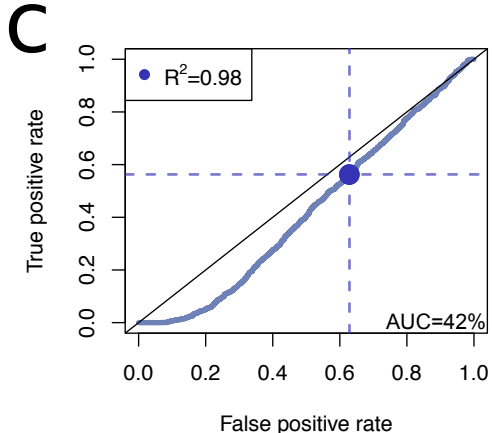
b

stochastic neutral model $\not\Rightarrow M(f) \propto \frac{1}{f}$

stochastic neutral model $\not\Leftarrow M(f) \propto \frac{1}{f}$



c



d

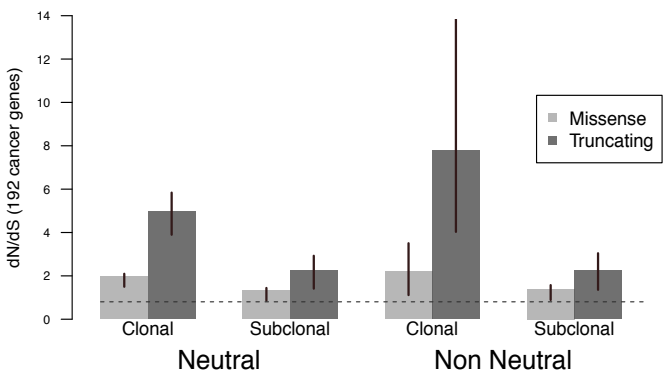


Figure 1 legend

(a) Neutrality calls in simulations of tumor growth with subclonal expansion underlying selective sweeps. The tree topology being modelled is represented on the right together with the parameters of the neutral evolution equations for the two subpopulations of cells (**Supplementary Methods**). The subclone's fraction (subclone %) increases with its selective advantage $adv_{subclone}$. We vary the $\lambda=1+adv_{subclone}$ and μ parameters of the subclone along a grid. Simulations are defined as true non-neutral (light blue) or false neutral (dark blue) when the growing subclone has expanded sufficiently to be detectable and the sweep is not complete, i.e. $10\% \leq$ subclone $\% \leq 90\%$, otherwise the subclone is considered beyond detection (light green). Non-neutral call: $R^2 < 0.98$; neutral call: $R^2 \geq 0.98$. **(b) As (a), using the Gillespie algorithm to simulate branching processes**[6]. Simulations leading to subclones beyond detection are either called neutral (light green) or non-neutral (dark green). Because of the stochastic nature of branching processes, different subclone % values are obtained across simulations from the same $adv_{subclone}$ values. For five increasing $adv_{subclone}$ values, we report median ± mad of the subclone % across the simulations. **(c) Summary ROC curve for the neutral vs. non-neutral classification based on the R^2 values in 1,919 non-neutral simulations from (b) and 1,919 simulations of neutral tumors.** The false positive rate and the true positive rate are highlighted for $R^2=0.98$ used by Williams *et al*. **(d) dN/dS analysis.** dN/dS ratios and confidence intervals for (sub)clonal mutations in TCGA tumors categorized into neutral and non-neutral groups. Ratios for missense and truncating mutations are given. dN/dS > 1 indicates positive selection.

References

1. Greaves, M. & Maley, C. C. CLONAL EVOLUTION IN CANCER. *Nature* **481,** 306–313 (2012).
2. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2016).
3. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
4. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48,** 238–244 (2016).
5. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
6. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Comput. Biol.* **12,** e1004731 (2016).
7. Martincorena, I. *et al.* Universal Patterns Of Selection In Cancer And Somatic Tissues. *bioRxiv* 132324 (2017). doi:10.1101/132324
8. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354,** 618–622 (2016).

Acknowledgments

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). MT is a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 747852-SIOMICS). PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. IM is funded by a Cancer Research UK Career Development Fellowship (C57387/A21777). This work was supported by grant 1U24CA210957 to PTS. FM would like to acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. Parts of this work were funded by CRUK core grant C14303/A17197. Parts of the results published here are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Members of the PCAWG Evolution and Heterogeneity Working Group

Stefan C. Dentro[1,2,3,*], Ignaty Leshchiner[4,*], Moritz Gerstung[5,*], Clemency Jolly[1,*], Kerstin Haase[1,*], Jeff Wintersinger[6,*], Pavana Anur[7], Rameen Beroukhim[4], Paul C. Boutros[6,8], David D. Bowtell[9,10], Peter J. Campbell[2], Elizabeth L. Christie[9], Marek Cmero[11], Yupeng Cun[12], Kevin Dawson[2], Jonas Demeulemeester[1,13], Amit Deshwar[6], Nilgun Donmez[14], Roland Eils[15,16], Yu Fan[17], Matthew Fittall[1], Dale W. Garsed[9], Gad Getz[4], Santiago Gonzalez[5], Gavin Ha[4], Marcin Imielinski[18,19], Yuan Ji[20,21], Kortine Kleinheinz[15,16], Juhee Lee[22], Henry Lee-Six[2], Dimitri G. Livitz[4], Geoff Macintyre[23], Salem Malikic[14], Florian Markowetz[23], Inigo Martincorena[2], Thomas J. Mitchell[2,24], Ville Mustonen[25], Layla Oesper[26], Martin Peifer[12], Myron Peto[7], Benjamin J. Raphael[27], Daniel Rosebrock[4], Yulia Rubanova[6], S. Cenk Sahinalp[28], Adriana Salcedo[8], Matthias Schlesner[15], Steve Schumacher[4], Subhajit Sengupta[20], Lincoln D. Stein[8], Maxime Tarabichi[1], Ignacio Vázquez-García[2,24], Shankar Vembu[6], Wenyi Wang[17], David A. Wheeler[29], Tsun-Po Yang[12], Xiaotong Yao[18,19], Fouad Yousif[8], Kaixian Yu[17], Ke Yuan[23,30], Hongtu Zhu[17], Quaid D. Morris[6,#], Paul T. Spellman[7,#], David C. Wedge[3,#], Peter Van Loo[1,13,#]

[1]The Francis Crick Institute, London NW1 1AT, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; [3]Big Data Institute, University of Oxford, Oxford OX3 7BN, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom; [6]University of Toronto, Toronto, ON M5S 3E1, Canada; [7]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97231, USA; [8]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; [9]Peter MacCallum Cancer Centre, Melbourne, VIC 3052, Australia; [10]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [11]University of Melbourne, Melbourne, VIC 3010, Australia; [12]University of Cologne, 50931 Cologne, Germany; [13]Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium; [14]Simon Fraser University, Burnaby, BC V5A1S6, Canada; [15]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; [16]Heidelberg University, 69120 Heidelberg, Germany; [17]The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [18]Weill Cornell Medicine, New York, NY 10065, USA; [19]New York Genome Center, New York, NY 10013, USA; [20]NorthShore University HealthSystem, Evanston, IL 60201, USA; [21]The University of Chicago, Chicago, IL 60637, USA; [22]University of California Santa Cruz, Santa Cruz, CA 95064, USA; [23]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, United Kingdom; [24]University of Cambridge, Cambridge CB2 0QQ, United Kingdom; [25]University of

Helsinki, 00014 Helsinki, Finland; [26]Carleton College, Northfield, MN 55057, USA; [27]Princeton University, Princeton, NJ 08540, USA; [28]Indiana University, Bloomington, IN 47405, USA; [29]Baylor College of Medicine, Houston, TX 77030, USA; [30]University of Glasgow, Glasgow G12 8RZ, United Kingdom.

*: These authors contributed equally

[#]: These authors jointly directed the work

Neutral tumor evolution? - Methods

Outline

First, we describe the two tumor growth models that were used. The first one is based on the deterministic continuous model presented by Williams *et al.*[1]. The second one is based on a branching process, a commonly-used discrete and fully stochastic growth model. We next explain how, using these two models, we can simulate variant allele fractions encountered in tumor sequencing studies. We describe our implementation of the approach by Williams *et al.*[1] to infer the most likely evolutionary path after the emergence of the most recent common ancestor (MRCA), i.e. neutral *vs*. non-neutral evolution. Finally, using real data from The Cancer Genome Atlas, we compare neutrality calls to results of dN/dS analysis, an independent and well-established approach to detect selection. We further describe the availability of the code as a tarball containing R and Java scripts and a Java runnable jar file called via one of the R scripts.

Simulations – continuous deterministic models

The deterministic equations described in Williams *et al.*[1] relate the number of cells in a tissue growing exponentially, N,

$$N(t) = 2^{\lambda\beta t}$$

and the cumulative number of mutations, M:

$$M(t) = \mu \int_0^t 2^{\lambda\beta t'} dt' = \frac{\mu}{\lambda\beta \ln(2)} \left(2^{\lambda\beta t} - 1\right) = \frac{\mu}{\lambda\beta \ln(2)} (N(t) - 1) \qquad \text{(Eq. 1)}$$

at any given time $t \geq 0$, where $\lambda > 0$ is the division rate per unit of time, $\beta \geq 0$ is the unitless "effective" division fraction, i.e. the fraction of divisions in which both daughter cells survive ($\beta = 1$ for no cell death, $\beta < 1$ to model cell death), and $\mu > 0$ is the mutation rate per cell division.

We have used these continuous deterministic models to simulate tumor growth *in silico* and followed each mutation and its corresponding variant cell fraction. To derive the cell fractions, we follow the progeny of the mother cell within which each mutation occurred.

Assume that the MRCA appears at time t_1, with division coefficient β_1, division rate λ_1, and mutation rate μ_1. To model a selective sweep within the cell

population spawned from the MRCA, we assume that at time $t_2 > t_1$, a subclone is initiated with division coefficient β_2, division rate λ_2, and mutation rate μ_2.

There is positive selection when $\lambda_2\beta_2 > \lambda_1\beta_1$. At time t the number of cells spawned from the MRCA but not part of the subclone (i.e. the cells with parameters β_1, λ_1, μ_1; further referred to as the MRCA lineage) is

$$N_1(t) = 2^{\lambda_1\beta_1(t-t_1)} - 2^{\lambda_1\beta_1(t-t_2)}$$

where the second term is omitted when $t < t_2$. Similarly, the number of cells at time t from the subclonal lineage (i.e. with parameters β_2, λ_2, μ_2) is

$$N_2(t) = 2^{\lambda_2\beta_2(t-t_2)}$$

when $t>t_2$ and $N_2(t)=0$ otherwise. The total cell count at time t is

$$N(t) = N_1(t) + N_2(t).$$

The tumor growth simulation is terminated at time $T > t_2$ and we derive the distribution at time T of the cell fractions for all mutations in the tumor.

Following the number of mutations and their cell fraction

Because the equations are continuous, they can lead to non-integer numbers of mutations and divisions. Hence, rather than deriving the number of mutations and their allele frequencies f at discrete time points, we model divisions in continuous time. We assess the number of additional mutations that have been added in fixed small time intervals of length dt. From Eq. (1) we find that the number of additional mutations occurring in the time interval $[t, t+dt]$ within a population of cells from the same lineage (i.e. parameters β, division rate λ, and mutation rate μ) is:

$$M(t+dt) - M(t) = \mu \frac{1}{\lambda\beta \ln(2)} (N(t+dt) - N(t))$$

For a mutation occurring at time t we may compute the variant cell fraction at time T. If the mutation occurred in a cell from the MRCA lineage that was not inherited by the subclone-initiating cell, then the variant cell fraction is

$$f_1(t) = \frac{2^{\lambda_1\beta_1(T-t)}}{N(T)}$$

If the mutation occurred in the subclone, then the variant cell fraction is

$$f_2(t) = \frac{2^{\lambda_2\beta_2(T-t)}}{N(T)}$$

Finally, if the mutation occurred in an ancestor cell of the subclone-initiating cell, then the variant cell fraction is

$$f_{12}(t) = \frac{2^{\lambda_1\beta_1(T-t)} - 2^{\lambda_1\beta_1(T-t_2)} + 2^{\lambda_2\beta_2(T-t_2)}}{N(T)}$$

Alternatively, we may calculate variant cell fractions in two steps, first determining the variant cell fraction of a mutation within the subpopulation of cells from the same lineage, and then scaling the variant cell fraction by the size of that subpopulation relative to the total cell population.

Setting the parameters for the grid of simulations

In each of our simulations the subclone growing under selective advantage appears at the 11th generation and the tumor is sampled at the 40th generation with a virtual purity of 100%. The number of initial clonal mutations μ_0 is not part of these models, and we arbitrarily set $\mu_0=\mu_2$. We fix the following parameters: clonal mutation rate $\mu_1 = 16$, clonal division rate $\lambda_1 = 1$, clonal division efficiency $\beta_1 = 0.4$, subclonal $\beta_2 = 0.4$. The depth of sequencing of the variants cov ~ Pois(10,000) to approach the theoretical distribution and the alternate read counts ~ Bin(cov, *f/2*), where f is the variant allele frequency derived from the model (see section on simulating tumor variant allele frequencies from sequencing data). We explore the results of the neutrality calls for a grid of parameter values

$$\mu_2 = \left\lceil (2^{0.5n})_{n\in\{2,3,\ldots,20\}} - 0.5 \right\rceil$$

and

$$adv_{subclone} = (0.01n)_{n\in\{0,1,2,\ldots,80\}},$$

where

$$adv_{subclone} = \lambda_2 - \lambda_1.$$

Simulations – fully stochastic models

To model stochastic discrete tumor growth, we used branching processes with the Gillespie algorithm[3]. These simulated tumors grow under asynchronous division, with zero or one subclone.

This was coded in Java. Each cell is a Java object and has four attributes: a Boolean value reporting whether the cell is alive or dead; an integer for the average number of mutations per division; an integer with mother cell ID; and an *ArrayList* of all *MutationSets* inherited from the mother cell. *MutationSet* is another class, for which each object contains one integer for the mother cell ID and one integer for the number of mutations within them. The constructor of *MutationSet* takes the mutation

rate of the mother cell as average number of events per interval of a Poisson distribution to draw the number of mutations.

Starting with an *ArrayList* of one tumor initiating cell, for each of 2^{20} cell division events, one cell is picked randomly from the living cells and either dies with probability P(*cell death*) or divides into two daughter cells with probability P(*division*) = 1 - P(*cell death*), akin to the Gillespie algorithm.

In our simulations, the subclone appears at the $2^{8\,th}$ division (~8^{th} generation) by changing the division rate value of one of the cells, and the tumor is sampled at the $2^{20\,th}$ division (~20^{th} generation). In these simulations, the number of mutations acquired at each cell division for each daughter cell is drawn from a Poisson distribution for the MRCA lineage $\mu \sim \text{Pois}(\mu_{\text{MRCA}})$ and the subclone lineage $\mu \sim \text{Pois}(\mu_{\text{subclone}})$.

The subclone is selected for division with probability

$$P(subclone\ divides) = \frac{(1 + adv_{subclone})N_{subclone}}{(1 + adv_{subclone})N_{subclone} + N_{MRCA}}$$

where $N_{subclone}$ and N_{MRCA} are the number of cells from the subclonal lineage and the MRCA lineage, respectively, and $adv_{subclone} > 0$ for positive selection and $adv_{subclone} = 0$ for neutral growth. The MRCA population will be selected for division with probability 1-P(*subclone divides*).

Within the selected clone, one cell is selected randomly for division with probability

$$P(cell\ divides) = \frac{1}{N}$$

where $N = N_{MRCA}$ if the cells belong to the MRCA lineage or $N = N_{subclone}$ if the cell belongs to the subclonal lineage.

With higher P(*cell death*), the first divisions are more likely to lead to the death of all cells and the tumor quickly stops growing. To limit this effect when cell death is high, we force the *D* first divisions to happen, i.e. P(*cell death*)=0 transiently until at least *2D* cells are alive.

Setting the parameters for the grid simulations

In our simulations, starting from one tumor initiating cell, for each of the 2^{20} cell division events, one cell is picked randomly and either dies with probability P(cell death) = 0.2 or divides into two daughter cells with probability P(division) = 1-P(cell death) = 0.8. The subclone appears at the $2^{8\,th}$ division (~8^{th} generation) and the

tumor is sampled at the 2^{20}-th division (~20^{th} generation). The ancestor clone's mutation rate $\mu \sim$ Pois(16). The average depth of coverage is 100X (see section on simulating tumor variant allele frequencies from sequencing data). In our simulations, D=6.

We explore a grid of values for

$$\mu_{subclone} = \left\lceil (2^{0.5n})_{n\in\{2,3,\ldots,20\}} - 0.5 \right\rceil$$

and

$$adv_{subclone} = (0.01n)_{n\in\{0,1,2,\ldots,100\}}.$$

This leads to 19*101=1,919 simulated tumors covering the grid.

Simulating tumor variant allele frequencies from sequencing data

Using the tumor growth models presented here, we can derive the exact number of mutations and their prevalence within a virtual tumor. These are taken as input to simulate the frequencies that would be observed in the sequencing reads from real tumor tissue.

In order to test the initial hypothesis, i.e. $M(f) \propto \frac{1}{f} \Leftrightarrow neutrality$, we start with the simplest models and assume: (i) the absence of non-tumor contaminant, (ii) 100% of the tumor cells are resected, and (iii) a fully diploid cancer genome.

Given exact cell fractions, *f,* of each mutation and an average sequencing coverage, *cov*, we draw for each individual mutation the total number of reads covering its genomic position *N* from a Poisson distribution $N \sim$ Pois(*cov*), and the alternate read counts *alt* ~ Bin(*N*,*f*/2), where *f*/2 is the allelic fraction for diploid regions. Finally, we generate variant calls by taking mutations with *alt* > 2 and derive the variant allelic fraction (VAF) of each variant $VAF = \frac{alt}{N}$. We then use the VAF distribution to call neutral and non-neutral tumors, as described by Williams *et al.*[1]

Calling neutral tumors

We followed the description by Williams *et al.*[1] to call neutral and non-neutral tumors based on the variant allele frequencies of their somatic single nucleotide variants. Tumors with less than 12 mutations with $0.12 \leq \text{VAF} \leq 0.24$ were removed. From the TCGA dataset, only tumors with a purity of at least 70%, as inferred by ASCAT[2], were analyzed.

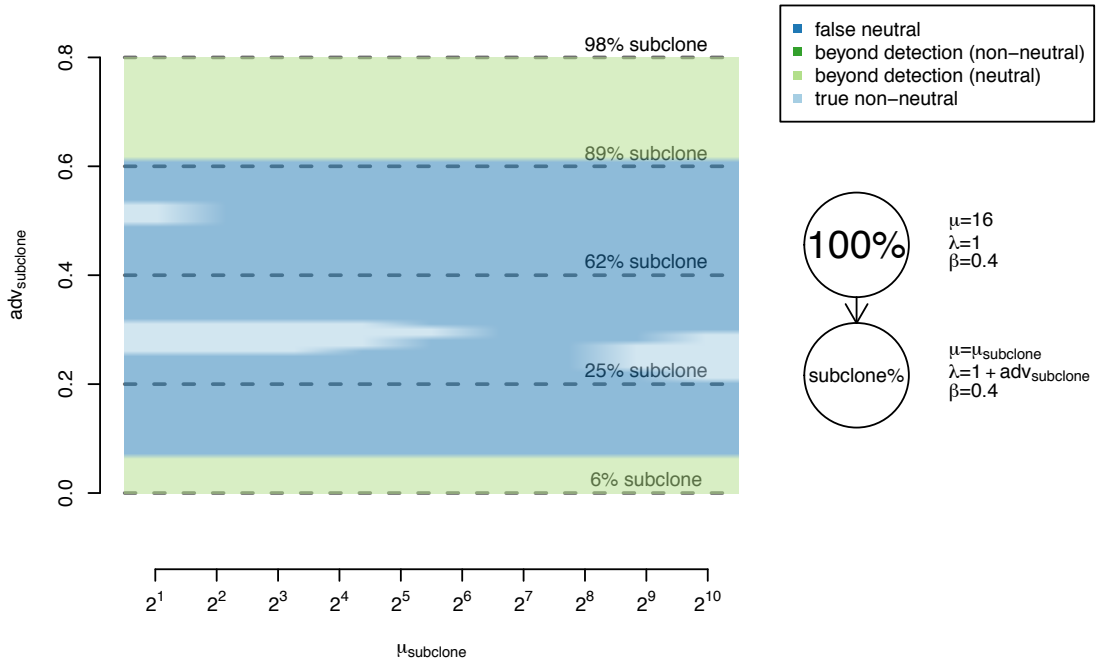
We calculated the explained variance (R^2) for linear regression models both with fixed intercept (intercept = 0) and without fixing the intercept, using the R commands:

> summary(lm(y~x+0,offset=rep(0,length(y))))$r.squared,

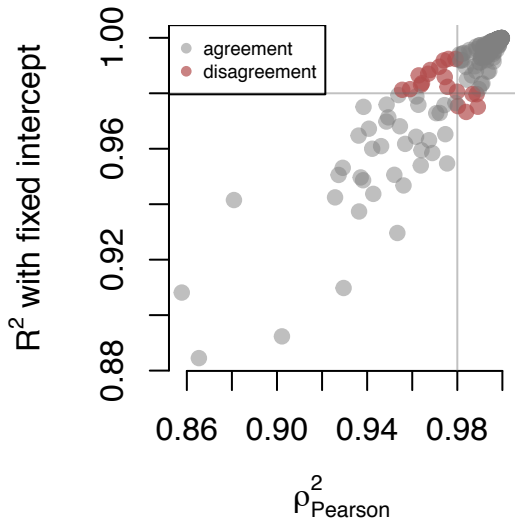
and

> cor(x,y)^2

respectively, where y is the cumulative number of mutations and x is the inverse allelic frequency minus the upper limit $x = \frac{1}{f} - \frac{1}{0.24}$. Results presented in the manuscript were obtained using a variable intercept. In **Supplementary Fig. 1**, we show the heat map of Figure 1a using a fixed intercept. Both methods show 97.5% agreement (**Supplementary Fig. 2**).



Supplementary figure 1. As reported in figure 1a using R^2 of a linear regression with fixed intercept = 0. The tree topology being modelled is represented on the right together with the parameters of the neutral evolution equations for the two subpopulations of cells. The subclone's fraction (subclone %) increases with its selective advantage $adv_{subclone}$. We vary the $\lambda=1+adv_{subclone}$ and μ parameters of the subclone along a grid. Simulations are defined as true non-neutral (light blue) or false neutral (dark blue) when the growing subclone is sizable enough to be detected and the sweep is not complete, i.e. 10% ≤subclone % ≤ 90%, otherwise the subclone is considered beyond detection (light green). Non-neutral call: $R^2 < 0.98$; neutral call: $R^2 \geq 0.98$.



Supplementary figure 2. R^2 values for the same simulations as in Supplementary figure 2, with variable and fixed intercept, showing an agreement of 97.5% on the neutral calls. The x-axis represents R^2 values (squared Pearson's correlation coefficients) for the linear regression between *M(f)* and *f* for the simulations in **Supplementary Fig. 1**. The y-axis represents R^2 values with fixed intercept = 0. Neutral calls, made if $R^2 \geq 0.98$, agree for 97.5% of these simulations (grey) and disagree for 2.5% of them (red).

ROC and area under the curve

Using fully stochastic branching processes, we simulated 1,919 non-neutral tumors and 1,919 neutral tumors and derived the R^2 values of the linear fit between the cumulative number of mutations and their inverse variant allelic fraction (VAF) within $0.12 \leq VAF \leq 0.24$. We then plotted the ROC using the R package ROCR version 1.0-7 and calculated the false positive rate and the true positive rate assuming the $R^2 = 0.98$ threshold used by Williams *et al.*[1]

Detection of selection in neutral and non-neutral tumors - dN/dS

Dataset

We ran our analyses on the data from The Cancer Genome Atlas, using CaVeMan[4,5] single nucleotide variant calls, and ASCAT[2] copy number calls, as described by Martincorena *et al.*[6]

Grouping variants into clonal and subclonal categories

To classify variants as clonal or subclonal, we used a one-sided proportion test to assess whether the alternate and total read counts of each variant were compatible

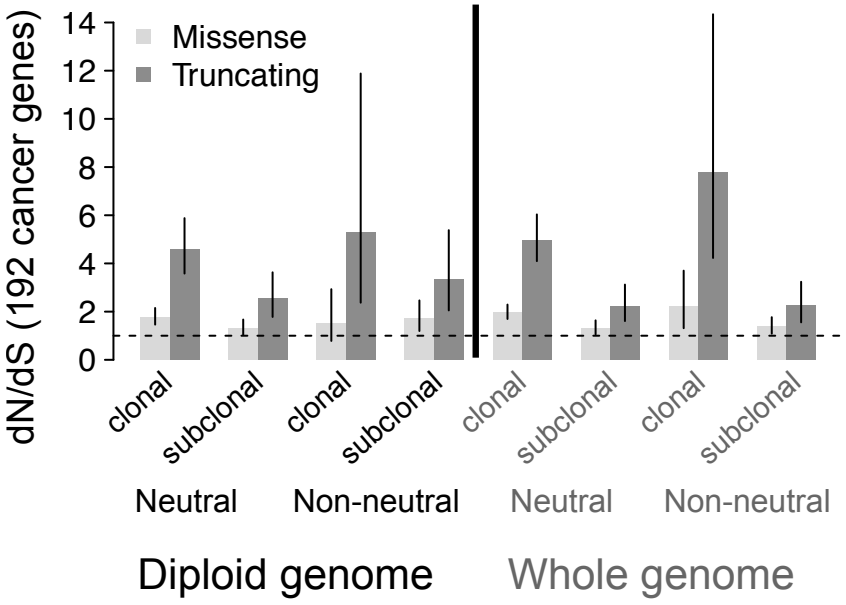
with its clonality, given its underlying number of DNA copies, and the overall tumor purity. This method is previously described in Alexandrov *et al.*[7]

dN/dS analysis

We performed dN/dS analysis to detect positive or negative selection of non-synonymous variants, as described by Martincorena *et al.*[6] We ran dN/dS separately on clonal and subclonal mutations and separately in the neutral and non-neutral tumors.

Effect of copy number

We repeated the analyses after selecting only variants that fall within diploid regions, i.e. 1 copy of allele A and 1 copy of allele B according to ASCAT[2], to show that the results were not induced by unreliable neutral calls, which could have resulted from the distortion of allele frequencies by copy number changes (**Supplementary Fig. 3**).



Supplementary Figure 3. dN/dS ratios on all mutations vs. mutations in diploid regions only, are shown for both missense and truncating mutations.

Code reproducibility and availability

Analyses and figures were generated using R version 3.1.3. The branching processes are coded in Java. The code for simulations is available as a tarball (included within this submission) with R scripts for the deterministic simulations and for deriving the figures, and a Java runnable jar file for generating variant fractions from the branching processes together with the associated Java source code.

References

1. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48,** 238–244 (2016).
2. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107,** 16910–16915 (2010).
3. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Comput. Biol.* **12,** e1004731 (2016).
4. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469,** 539–542 (2011).
5. Jones *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinforma.* **56,** 15.10.1-15.10.18 (2016).
6. Martincorena, I. *et al.* Universal Patterns Of Selection In Cancer And Somatic Tissues. *bioRxiv* 132324 (2017). doi:10.1101/132324
7. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354,** 618–622 (2016).