nature genetics

Identification of neutral tumor evolution across cancer types

Marc J Williams[1–3,6], Benjamin Werner[4,6], Chris P Barnes[2,5], Trevor A Graham[1] & Andrea Sottoriva[4]

Despite extraordinary efforts to profile cancer genomes, interpreting the vast amount of genomic data in the light of cancer evolution remains challenging. Here we demonstrate that neutral tumor evolution results in a power-law distribution of the mutant allele frequencies reported by next-generation sequencing of tumor bulk samples. We find that the neutral power law fits with high precision 323 of 904 cancers from 14 types and from different cohorts. In malignancies identified as evolving neutrally, all clonal selection seemingly occurred before the onset of cancer growth and not in later-arising subclones, resulting in numerous passenger mutations that are responsible for intratumoral heterogeneity. Reanalyzing cancer sequencing data within the neutral framework allowed the measurement, in each patient, of both the *in vivo* mutation rate and the order and timing of mutations. This result provides a new way to interpret existing cancer genomic data and to discriminate between functional and non-functional intratumoral heterogeneity.

Unraveling the evolutionary history of a tumor is clinically valuable, as prognosis depends on the future course of the evolutionary process and therapeutic response is mostly determined by the evolution of resistant subpopulations[1,2]. In humans, the details of tumor evolution have remained largely uncharacterized, as longitudinal measurements are impractical and studies are complicated by between-patient variation[3] and intratumoral heterogeneity (ITH)[4,5]. Several recent studies have begun tackling this complexity[6], highlighting patterns of convergent evolution[7], punctuated dynamics[8] and intricate interactions between cancer cell populations[9]. However, the lack of a rigorous theoretical framework able to make predictions on existing data[10] means that results from cancer genomic profiling studies are often difficult to interpret. For example, how much of the detected ITH is actually functional is largely unknown, also because a rigorous 'null model' of genomic heterogeneity is lacking. In particular, interpreting the mutant allele frequency distribution reported by next-generation sequencing is problematic because of the absence of a formal model linking tumor evolution to the observed data.

Here we show that the subclonal mutant allele frequencies of a substantial proportion of cancers of different types and from different cohorts precisely follow a simple power-law distribution predicted by neutral growth. In neutral cancers, all tumor-driving alterations responsible for cancer expansion appear to have been present in the first malignant cell, and subsequent tumor evolution was effectively neutral. We demonstrate that, under neutral growth, fundamental parameters describing cancer evolution that have been thus far inaccessible in human tumors, such as the mutation rate and the mutational timeline, become measurable. Notably, this approach also allows the identification of non-neutral malignancies, in which ongoing clonal selection and adaption to microenvironmental niches may have a strong role during cancer growth.

RESULTS

Neutral cancer growth

Recently, we showed that colorectal cancers (CRCs) often grow as a single expansion, populated by a large number of intermixed subclones[11]. Consequently, we expect that, after malignant transformation, individual subclones with distinct mutational patterns will grow at similar rates, coexisting within the tumor for long periods of time without overtaking one another, as a result of the lack of stringent selection. Moreover, only a handful of recurrent driver alterations have been identified in CRC[12], and these are reported to be ubiquitous in multiregion sampling[11] and stable during cancer progression[13], indicating that they were all present in the 'first' cancer cell and that subsequent clonal outgrowths are relatively rare. Therefore, we hypothesized that cancer evolution may often be dominated by neutral evolutionary dynamics.

The dynamics of neutral evolutionary processes have been widely studied in the context of molecular evolution and population genetics[14–16] as well as in mouse models of cancer[17]. However, the widely held presumption that subclone dynamics in human cancers are dominated by strong selection has meant that these ideas have been neglected in current studies of cancer evolution.

Motivated by this, here we present a theoretical model describing the expected pattern of subclonal mutations within a tumor that is evolving according to neutral evolutionary dynamics. The model postulates that, after the accumulation of a 'full house' of genomic changes that initiate cancer growth, some cancers expand neutrally, generating a large number of passenger mutations that are responsible for the extensive and common ITH. The parameter-free model is applicable to next-generation sequencing data from any solid cancer. Here we present the model and, by applying it to large, preexisting

[1]Evolution and Cancer Laboratory, Barts Cancer Institute, Queen Mary University of London, London, UK. [2]Department of Cell and Developmental Biology, University College London, London, UK. [3]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK. [4]Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK. [5]Department of Genetics, Evolution and Environment, University College London, London, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to T.A.G. (t.graham@qmul.ac.uk) or A.S. (andrea.sottoriva@icr.ac.uk).

Received 18 August 2015; accepted 18 December 2015; published online 18 January 2016; doi:10.1038/ng.3489

cancer genomic data sets, determine which tumors are consistent with neutral growth. When the model applies, we measure new tumor characteristics directly from the patient's data.

Model derivation

A cancer is founded by a single cell that has already acquired a substantial mutation burden[3]: these 'precancer' mutations will be borne by every cell in the growing cancer and so become 'public', or clonal. Mutations that occur within different cell lineages remain 'private', or subclonal, in an expanding malignancy under the absence of strong selection. Here we focus on subclonal mutations, as they contain information on the dynamics of cancer growth. We denote the number of cancer cells at time t as $N(t)$, with cells dividing at rate λ per unit time. During a cell division, somatic mutations occur at rate μ. If we consider an average number of π chromosome sets in a cancer cell (the ploidy of the cell), we can calculate the expected number of new mutations per time interval as

$$\frac{dM}{dt} = \mu\pi\lambda N(t) \tag{1}$$

Solving this requires integrating over the growth function $N(t)$ in some time interval $[t_0,t]$

$$M(t) = \mu\pi\lambda \int_{t_0}^{t} N(t)dt \tag{2}$$

Because not all cell divisions may be successful in generating two surviving lineages, as a result of cell death or differentiation, we introduce the fraction β of 'effective' cell divisions in which both resulting lineages survive. In the case of exponential growth, the mean number of tumor cells as a function of time is therefore

$$N(t) = e^{\lambda\beta t} \tag{3}$$

Substituting into equation (2) gives the explicit solution

$$M(t) = \frac{\mu\pi}{\beta}\left(e^{\lambda\beta t} - e^{\lambda\beta t_0}\right) \tag{4}$$

This equation describes the total number of subclonal mutations that accumulate within a growing tumor in the time interval $[t_0,t]$. We note that, for $t_0 = 0$, equation (4) corresponds to the Luria-Delbrück model, which describes mutation accumulation in bacteria[18]. In our case, this equation is of limited use, as none of the parameters μ, λ, β or the age of the tumor t can be measured directly in humans. However, we do know that, for a new mutation occurring at any time t, its allelic frequency (relative fraction) f must be the inverse of the number of alleles in the population

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda\beta t}} \tag{5}$$

For example, if a new mutation arises in a tumor of 100 cells, it will comprise a cellular fraction of 1/100. In the absence of clonal selection (or, indeed, substantial genetic drift), the allelic frequency of a mutation will remain constant during expansion, as all cells, with and without this mutation, grow at the same rate. In the given example, after one generation has elapsed, we will have two cells with that particular mutation but a total of 200 tumor cells, again yielding a fraction of 1/100. This implies that, in the neutral case, tumor age t and mutation frequency f are interchangeable. For example, $t_0 = 0$ in a diploid tumor ($\pi = 2$) corresponds to $f_{max} = 0.5$ (the expected allelic frequency of clonal variants), where

$$f_{max} = \frac{1}{\pi e^{\lambda\beta t_0}} \tag{6}$$

Substituting t for f in equation (4) gives an expression for the cumulative number of mutations in the tumor per frequency $M(f)$

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \tag{7}$$

thus converging to the solution for expanding populations under neutrality obtained using other approaches[19–22]. Critically, the distribution $M(f)$ is naturally provided by next-generation sequencing data from the bulk sequencing of tumor biopsies and resections, against which the model can be tested. The model predicts that mutations arising during the neutral expansion of a cancer accumulate following a $1/f$ power-law distribution. In other words, when neutral evolution occurs in a tumor, the number of subclonal mutations detected should accumulate linearly with the inverse of their frequency. The $1/f$ noise, or 'pink noise', is common in nature and is found in several physical, biological and economic systems[23].

Notably, the coefficient $\mu_e = \mu/\beta$ is the mutation rate per effective cell division and corresponds to the easily measurable slope of $M(f)$. This model therefore provides a straightforward, parameter-free method to measure the *in vivo* mutation rate in a patient's tumor using a single next-generation sequencing sample. We note that the results do not depend on the identity of the alterations considered, as any genomic alteration (mutation, copy number change or epigenetic modification) that changes the dynamics of tumor growth (for example, any alteration that is clonally selected) would likely result in deviation from the neutral $1/f$ power law by causing an over- or under-representation of the alleles in that clone. Hence, this analysis uses single-nucleotide variants (SNVs) as 'barcodes' to follow clone growth. Stochastic simulations of neutral tumor growth support the analytical solution in equation (7) (Online Methods).

Identification of neutrality in colorectal cancer evolution

A typical allelic frequency distribution of mutations in a tumor measured by whole-exome sequencing is shown in **Figure 1a** (data from ref. 11). Considering tumor purity and aneuploidy, mutations with high allelic frequency (>0.25) are likely to be public (clonal), whereas all others are likely to be subclonal. The same data can be represented as the cumulative distribution $M(f)$ of subclonal mutations, as in equation (7) (**Fig. 1b**). Remarkably, as represented by the high goodness-of-fit measure R^2, these data precisely follow the distribution predicted by the model, indicating that this tumor grew under neutral evolutionary dynamics.

We next considered our cohort of seven multiple-sample CRCs[11] and 101 colon adenocarcinomas[12] from The Cancer Genome Atlas (TCGA) selected for high tumor purity (≥70%) that underwent whole-exome sequencing (Online Methods). The latter set was separated into tumors characterized by chromosomal instability (CIN) and tumors with microsatellite instability (MSI). The power law was remarkably well supported in both these cohorts, with 38 of 108 (35.1%) of the cases reporting a high R^2 value ≥0.98 (**Fig. 1c**). These results confirm that, in a large proportion of colon cancers, within-tumor clonal dynamics are not dominated by strong selection but rather follow neutral evolution. In particular, a larger proportion of cancers with CIN evolved neutrally (31/82; 37.8%) than did cancers with MSI (3/19; 15.7%) (**Fig. 1c**), possibly because the latter acquired so many new

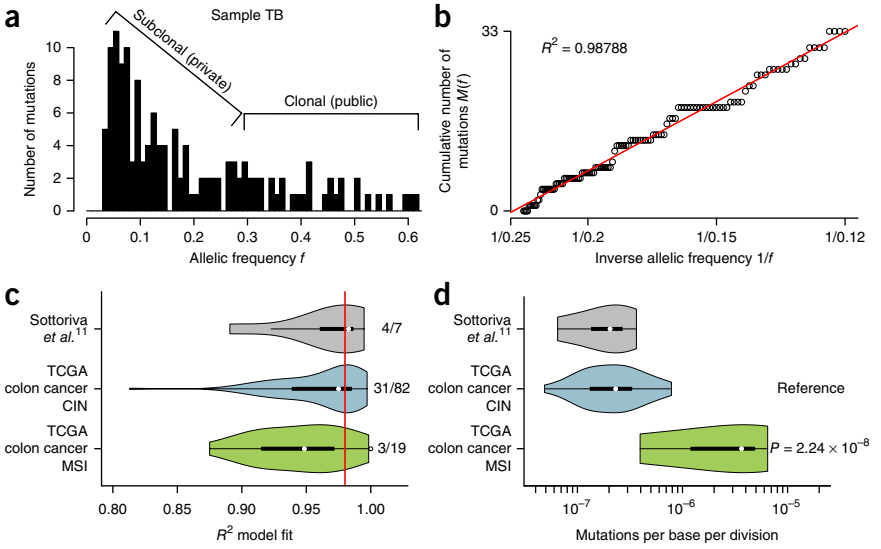

Figure 1 Neutral evolution is common in colon cancer and allows measurement of the mutation rate in each tumor. (**a**) The output of next-generation sequencing, such as whole-exome sequencing, can be summarized as a histogram of mutant allele frequencies, here for sample TB. Considering purity and ploidy, mutations with relatively high frequency (>0.25) are likely to be clonal (public), whereas low-frequency mutations capture the tumor subclonal architecture. (**b**) The same data can be represented as the cumulative distribution $M(f)$ of subclonal mutations. This distribution was found to be linear with $1/f$, precisely as predicted by the neutral model. (**c**) The R^2 goodness-of-fit measure for our CRC cohort ($n = 7$) and the TCGA colon cancer cohort ($n = 101$) grouped as having CIN or MSI confirmed that neutral evolution is common (38/108; 35.1% of samples with $R^2 \geq 0.98$). The red line indicates the $R^2 = 0.98$ threshold for discriminating neutral from non-neutral tumors. (**d**) Measurements of the mutation rate showed that the groups with CIN had a median mutation rate of $\mu_e = 2.31 \times 10^{-7}$, whereas tumors with MSI had a 15-fold higher rate (median $\mu_e = 3.65 \times 10^{-6}$; F test, $P = 2.24 \times 10^{-8}$), as predicted on the basis of their deficiency in DNA mismatch repair.

mutations that some were likely under strong selection. Because $M(f)$ is a monotonic growing function, a stringent threshold of $R^2 \geq 0.98$ was chosen to prevent overcalling neutrality, but we note that we may have therefore misclassified some tumors as non-neutral because of limited sequencing depth or low mutation burden. R^2 values were independent of the mean coverage of mutations, the total number of mutations in the sample and the number of mutations within the model range (Online Methods). See **Supplementary Data Set 1** for a summary of the TCGA data used.

Measurement of the mutation rate in colorectal cancer

Estimating the per-base mutation rate μ per division in human malignancies is challenging because direct measurements are not possible. Previous estimates critically depend on assumptions about the duration of the cell cycle and the growth rate λ, as well as on the total mutation burden of the cancer[24–26]. However, accurate measurement of all mutations within a cancer, including heterogeneous subclonal variants, is technically unfeasible because most mutations are present in very small numbers of cells[4]. With our approach, it is possible to circumvent this issue by measuring the rate of accumulation of subclonal mutations represented by the slope of $M(f)$. In the case of neutral evolution, this can be done in principle within any (subclonal) frequency range, without the need of detecting extremely rare mutations. We estimated the mutation rate in all samples with $R^2 \geq 0.98$ (**Fig. 1d**) and found that it was more than 15-fold higher in the MSI group (median $\mu_e = 3.65 \times 10^{-6}$) than in the CIN group (median $\mu_e = 2.31 \times 10^{-7}$; F test, $P = 2.24 \times 10^{-8}$) and our cohort of CRCs (median $\mu_e = 2.07 \times 10^{-7}$), of which all but one tumor had CIN[11]. Different mutation types (for example, transitions and transversions) are caused by particular mutational processes[27] and thus likely occur at different rates; accordingly, we found that C>T mutations occurred at median rate $\mu_{e,C>T} = 2.19 \times 10^{-7}$, a rate nearly tenfold higher than that for any other type of mutation (F test, $P = 3.13 \times 10^{-3}$; **Supplementary Fig. 1a**). We stratified according to CIN versus MSI status and found that the mutation rate for each mutation type reflected the overall mutation rate for the group (**Supplementary Fig. 1b**). The variation in mutation rates within and between subgroups was remarkably in line with the variation in estimates of mutation burden in colon cancer[3]. We note that the mutation rate estimate is scaled by the (unknown) effective division rate β, which means for example that, if only one in 100 cell divisions leads to two surviving offspring ($\beta = 0.01$), then the mutation rate μ is 100 times lower than the effective rate μ_e reported. The mutation rates of non-neutral cases ($R^2 < 0.98$) cannot be estimated, as the model does not fit the dynamics of these tumors.

We examined the effect of copy number changes in the model by performing the analysis using only mutations in diploid regions and found highly similar proportions of neutral tumors and mutation rates (Online Methods and **Supplementary Fig. 2**). The validity of the variant calls was also corroborated by the consistency of the underlying mutational signature across a range of allelic frequencies; hence, the results are unlikely to have been influenced by sequencing errors (**Supplementary Fig. 3**).

Frequent selection events should induce a higher number of missense and nonsense mutations than expected by chance, whereas under neutrality we expect the same rate of silent and non-silent mutations. To test this, we contrasted the estimated rate of synonymous mutations (unlikely to be under selection) with the rate of missense and nonsense mutations (liable to experience selection). Although the latter were more common than the former, after adjustment for the number of potential synonymous and nonsynonymous sites in the exome, the two rates were equivalent (**Supplementary Fig. 4**), consistent with neutral evolution.

Neutral evolution in coding and noncoding regions

We next tested whether the signature of neutral evolution could be detected across the entire genome, not just in coding regions. To do this, we analyzed 78 gastric cancers from a recent study[28] subjected to high-depth whole-genome sequencing and preselected for high tumor purity (≥70%). The large number of mutations detected by whole-genome sequencing accumulated precisely as predicted by the model (example in **Fig. 2a**,**b**), indicating neutral evolution in 60 of 78 (76.9%) cases (**Fig. 2c**). A smaller proportion of tumors with MSI were neutral (3/10; 30%) than microsatellite-stable (MSS) tumors (57/68; 83.8%), in line with the observation in CRC. A tumor was consistently classified as neutral independently of whether all SNVs or only noncoding SNVs were used to perform the classification (**Fig. 2c**, Venn diagram), whereas, because of the limited number of mutations available in the exome coupled with the strict $R^2 = 0.98$ threshold to call neutrality, fewer tumors were identified as neutral using exonic

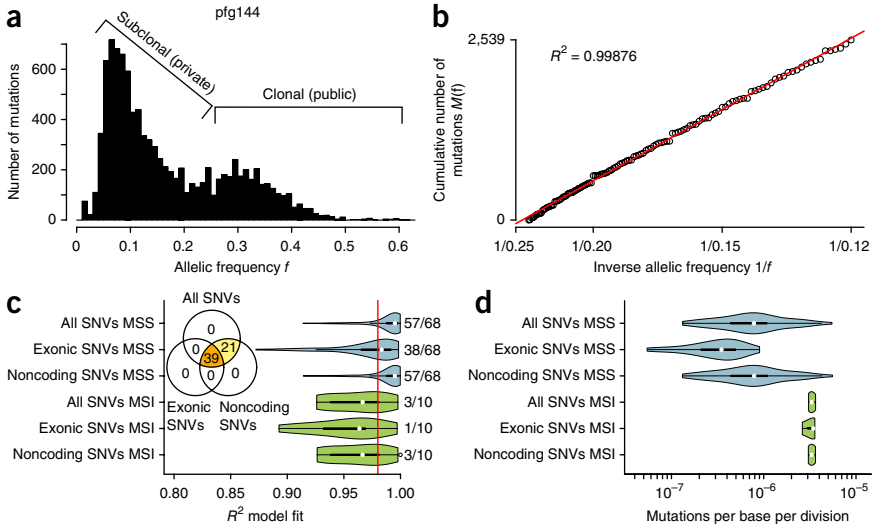

Figure 2 Neutral evolution across the whole genome of gastric cancers. (**a**) A large number of coding and noncoding mutations can be identified using whole-genome sequencing. (**b**) All detected mutations precisely accumulate as $1/f$ following the neutral model in this example. (**c**) Neutral evolution is very common in gastric cancer, with 60 of 78 (76.9%) samples showing goodness of fit for the neutral model $R^2 \geq 0.98$ (red line). This was consistent using all, exonic or noncoding subclonal mutations. The same tumors were identified as neutral by all three methods, although limitations in detecting neutrality were present when considering exonic mutations because of the limited number of variants. WGS, whole-genome sequencing. (**d**) Mutation rates were more than four times higher in MSI ($\mu_e = 3.30 \times 10^{-6}$) than in MSS ($\mu_e = 7.82 \times 10^{-7}$) tumors ($F$ test, $P = 1.35 \times 10^{-4}$), consistent with the underlying biology.

SNVs alone. Notably, every case was verified as neutral by at least two different variant sets. These results suggest that neutral evolution can be robustly assessed from mutations anywhere in the genome.

Mutation rate analysis of the neutrally evolved gastric cancers (**Fig. 2d**) showed that cancers with MSI had a more than fourfold higher mutation rate ($\mu_e = 3.30 \times 10^{-6}$) than MSS tumors ($\mu_e = 7.82 \times 10^{-7}$; F test, $P = 1.35 \times 10^{-4}$). The results were robust to copy number changes when the analysis was performed using only variants in diploid regions (**Supplementary Fig. 5**). The mutational signature of the variant calls for this cohort was also consistent across the frequency spectrum (**Supplementary Fig. 6**). Synonymous versus nonsynonymous mutation rates were also consistent with neutral evolution (**Supplementary Fig. 7**). See **Supplementary Data Set 2** for a summary of the data from Wang *et al.* used.

Neutral evolution across cancer types

We then applied the neutral model to a large pan-cancer cohort of 819 exome-sequenced cancers from 14 tumor types from the TCGA Consortium (which included the 101 colon cancers previously examined). All of these samples had been preselected for high tumor purity (≥70%). The fit of the model was remarkably good across cancer types (**Fig. 3a**), with 259 of 819 (31.6%) cases showing $R^2 \geq 0.98$. We found that neutral evolution seemed more prominent in some tumor types, such as stomach (validating the whole-genome sequencing analysis), lung, bladder, cervical and colon cancers. Other tumor types showed a consistently poorer fit, indicating that the clonal dynamics in these malignancies were typically not neutral, such as in renal cancer, melanoma, pancreatic cancer, thyroid cancer and glioblastoma. Consistent with these results, non-neutral renal carcinoma has been shown to display convergent evolution in spatially disparate tumor regions driven by strong selective forces[7], whereas the same phenomenon was not found in more neutral lung cancer[29,30]. Other tumor types displayed mixed dynamics, with some cases that were characterized by neutral evolution and some that were not. We note that a proportion of melanoma samples in this cohort are derived from regional metastases and not primary lesions, and this could potentially explain the lack of neutral dynamics observed.

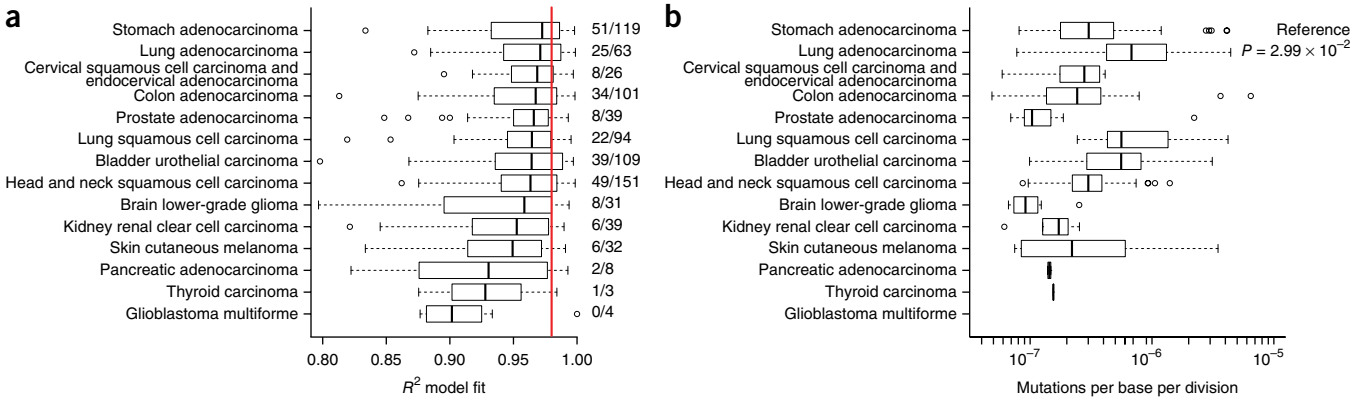

Figure 3 Neutral evolution and mutation rates across cancer types. (**a**) R^2 values from 819 cancers of 14 different types supported neutral evolution in a large proportion of cases (259/819; 31.6% with $R^2 \geq 0.98$; red line) and across different cancer types, particularly in stomach (validating the whole-genome sequencing analysis), lung, bladder, cervical and colon cancers. On the contrary, renal cancer, melanoma, pancreatic cancer, thyroid cancer and glioblastoma were characterized by non-neutral evolution. The other types displayed mixed dynamics. (**b**) The highest mutation rates were found in lung cancer. Lower rates were found in thyroid cancer, lower-grade glioma and prostate cancer. Box plots show the median, first and third quartiles (box), and the lowest data point that is 1.5 times the interquartile range (IQR) of the lower quartile and the highest data point still within 1.5 times the IQR of the upper quartile (whiskers) of the data; circles are outlying data points.

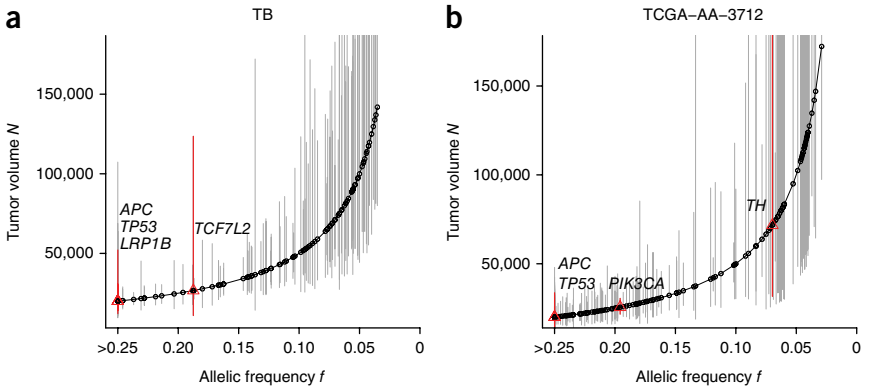

Figure 4 Reconstruction of the mutational timeline in two patients. The allelic frequency of a mutation within the tumor predicts the size of the tumor when the mutation occurred. (**a**,**b**) Deconvolution of the mutational timeline is illustrated for samples TB (**a**) and TCGA-AA-3712 (**b**). Whereas established CRC driver alterations (in *APC*, *KRAS* and *TP53*) seem to be present from the first malignant cell, several recurrent putative drivers not yet validated were present after malignant seeding, despite the underlying neutral dynamics. This suggests that some of these candidate alterations may not be fundamental drivers of growth in all cases. Confidence intervals were calculated using a binomial test on the number of variant reads versus the depth of coverage for each mutation.

Mutation rate analysis on the neutral cases showed differences of more than one order of magnitude between tumor types (**Fig. 3b**). The highest mutation rates were observed in lung adenocarcinoma (median $\mu_e = 6.79 \times 10^{-7}$) and lung squamous cell carcinoma (median $\mu_e = 5.61 \times 10^{-7}$), and the lowest rates were seen in low-grade glioma (median $\mu_e = 9.22 \times 10^{-8}$) and prostate cancer (median $\mu_e = 1.04 \times 10^{-7}$). We stratified the mutation rates into different mutational types (**Supplementary Fig. 8**) and found that C>A mutations occurred at a significantly higher rate in lung cancers (lung adenocarcinoma, $P = 2.72 \times 10^{-16}$; lung squamous cell carcinoma, $P = 3.90 \times 10^{-4}$), consistent with their causation by tobacco smoke[27]. C>T mutation rates were most consistent across cancer types, likely because of their association with replicative errors as opposed to being caused by a particular stochastically arising defect in DNA replication or repair[27].

These results demonstrate that within-tumor clonal dynamics can be neutral, and the classification of tumors on the basis of neutral versus non-neutral growth dynamics leads to new measurements of fundamental tumor biology. See **Supplementary Data Set 1** for a summary of the TCGA data used.

In silico validation of the neutral model

To assess the different inherent sources of noise in next-generation sequencing data (contamination from normal tissue, limited sequencing depth and tumor sampling), we designed a stochastic simulation of neutral growth that produced synthetic next-generation sequencing data from bulk samples (Online Methods). The simulations produced realistic-looking synthetic next-generation sequencing data (**Supplementary Fig. 9**) with minimal assumptions and under a range of different scenarios for tumor growth dynamics (variable low mutation rate and variable number of clonal mutations) and sources of assay noise (normal contamination in the sample, sequencing depth and detection limit). For each of these potentially confounding factors, we were able to fit our neutral model to the synthetic next-generation sequencing data and accurately recover both the underlying neutral dynamics and the mutation rate (**Supplementary Fig. 10**). We also validated the prediction that $M(f)$ would deviate from the neutral power law in the presence of emerging subclones with a higher fitness advantage (**Supplementary Fig. 11a**,**b**), as well as in the case of a mixture of subclones (as observed in ref. 31) emerging either by means of clonal expansions triggered by selection or segregating microenvironmental niches (**Supplementary Fig. 11c–f**). Variation in mutation rate between subclones also caused a deviation from neutrality (**Supplementary Fig. 11g,h**). These results support the reliability of the conservatively high R^2 threshold used to call neutrality.

Mutational timelines

Under neutral evolution, it is possible to estimate the size of the tumor when a mutation with frequency f arose using equation (5)

$$N(t) = \frac{1}{\pi f} \tag{8}$$

The decomposition of the mutational timeline for two illustrative cases—sample TB from ref. 11 and sample TCGA-AA-3712 from ref. 12—is shown in **Figure 4a**,**b**. Previous estimates of mutational timelines relied on cross-sectional data[32–35], which are compromised by the extensive heterogeneity, whereas multiregion profiling approaches are more accurate but are also more expensive and laborious[7,36,37]. Using our formal model of cancer evolution, the timeline information becomes accessible from routinely available genomic data. We found that classical CRC driver alterations, such as in the *APC*, *KRAS* and *TP53* genes, indeed seemed to be present in the first malignant cell (likely because they accumulated during previous neoplastic stages). This finding agrees with what we previously reported using single-gland mutational profiling where all these drivers, when present, were found in all glands[11]. However, we also found that, when we considered a more extended list of putative driver alterations, many occurred during the neutral phase of tumor growth, suggesting that the selective advantage conferred by a putative driver alteration may be context dependent, as demonstrated in a *Trp53* mouse model[38].

DISCUSSION

Understanding the evolutionary dynamics of subclones within human cancers is challenging because longitudinal observations are unfeasible and the genetic landscape of cancer is highly dynamic, leading to genomic data that are hard to interpret[39]. In particular, complex, nonlinear evolutionary trajectories have been observed, such as punctuated evolution and karyotypic chaos[8,39,40]. Here we have presented a formal law that predicts mutational patterns routinely reported in the next-generation sequencing of bulk cancer specimens. Our analysis of large independent cohorts using this framework shows that cancer growth often seems to be dominated by neutral evolutionary dynamics, an observation that is consistent across 14 cancer types. Under neutrality, the clonal structure of a tumor is expected to have a fractal topology characterized by self-similarity (**Fig. 5**). As the tumor grows, a large number of cell lineages are generated, and ITH therefore rapidly increases while the allele frequency of the new heterogeneous mutations quickly decreases because of expansion. This implies that sampling in different parts of the tree leads to the detection of distinct mutations that all show the same $1/f$ distribution. Clonal mutations

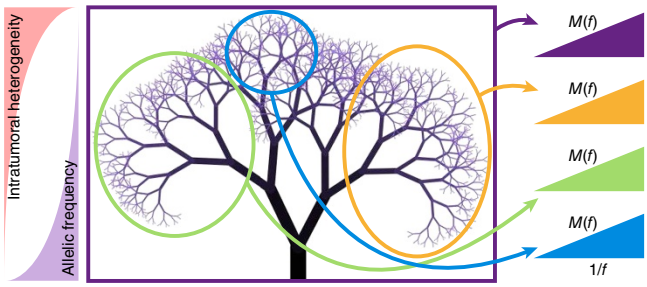

Figure 5 Neutral evolution and tumor phylogeny. After the accumulation of key genomic alterations, in neutral malignancies, cancer expansion is likely triggered by a single critical genomic event (the accumulation of a 'full house' of genomic changes) followed by neutral evolution that generates a large number of new mutations in ever smaller subclones. While tumor heterogeneity rapidly increases, the allele frequency of heterogeneous mutations decreases. In this context, the accumulation of mutations $M(f)$ follows a characteristic $1/f$ distribution. Moreover, the tumor phylogeny displays a characteristic fractal topology that is self-similar. Sampling in different regions of the phylogenetic tree exposes distinct mutations that show the same $1/f$ distribution. Clonal mutations in a sample (not considered in the model) arose in the most recent common ancestor of the sampled cells. Because of the large population of cells sampled using bulk sequencing, the majority of detected clonal mutations belong to the trunk of the tree and therefore are likely found in the first cancer cell. Deviations from the $1/f$ law indicate different dynamics from neutral growth.

found in a sample (not considered in the model) belong to the most recent common ancestor in the tree.

We note that some cancers were dominated by neutral evolution whereas others were not. In non-neutral tumors, strong selection, microenvironmental constraints and non–cell autonomous effects[41] may have a key role. Notably, this formalization represents the null model of cancer within-clone heterogeneity that can be used to identify cases in which complex, non-neutral dynamics occur and to discriminate between functional and non-functional ITH. Furthermore, we speculate that neutral evolutionary dynamics may be favored by the cellular architecture of the tumor (for example, glandular structures that limit the effects of selection) and/or the anatomical location of the malignancy (for example, a lumen versus a highly confined space), as well as the presence of potentially selective microenvironmental features of the tumor such as hypoxic regions. Despite the evidence for a lack of natural selection during malignant growth, eventual treatment is likely to change the rules of the game and strongly select for treatment-resistant clones[42]. Clones with driver alterations underlying treatment resistance that were not under selection during growth may expand as a result of new selective pressures introduced by therapy. The same may happen in the context of the purported evolutionary bottleneck preceding metastatic dissemination. Notably, this reasoning highlights how 'drivers' can only be defined within a context and so the same driver alteration can be neutral in a certain microenvironmental context (for example, in the absence of treatment) and not neutral in another (for example, during treatment). Moreover, we predict that, if a tumor is characterized by different microenvironmental niches but still presents as neutral, it is likely that adaptation will be driven by cancer cell plasticity rather than clonal selection. Cell plasticity is hard to study in cancer because it implies a change in the cell phenotype that is not caused by inheritable genomic variation. Thus, this phenomenon has been so far largely neglected in cancer. As neutrality can be used as the null model with which to identify clonal selection, this model facilitates the study of adaptation through plasticity directly in human malignancies. Furthermore, it is important to note that, because of the intrinsic detection limits of sequencing technologies, it is possible to explore only the early expansion of cancer clones (**Fig. 5**), and the dynamics of extremely small clones may remain undetected.

Notably, the realization that within-tumor clonal dynamics can be neutral means that the *in vivo* mutation rate per division and the mutational timeline—factors that have a key role in cancer evolution, progression and treatment resistance—can be measured directly from patient data. These measurements can be performed in a patient-specific manner and so may be useful for prognostication and the personalization of therapy. Recognizing that the growth of a neoplasm can often be dominated by neutral dynamics provides an analytically tractable and rigorous method to study cancer evolution and gain clinically relevant insight from commonly available genomic data.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Shibata, C. Curtis, S. Tavaré and R. Durrett for fruitful discussions. We would like to thank N. Andor (Stanford University) for supplying mutation calls for the TCGA data. We also thank V. Mustonen for useful suggestions.

A.S. is supported by The Chris Rokos Fellowship in Evolution and Cancer. B.W. is supported by the Geoffrey W. Lewis Post-Doctoral Training fellowship. This work was supported by the Wellcome Trust (105104/Z/14/Z). C.P.B. acknowledges funding from the Wellcome Trust through a Research Career Development Fellowship (097319/Z/11/Z). This work was supported by a Cancer Research UK Career Development Award to T.A.G. M.J.W. is supported by a UK Medical Research Council student fellowship.

This study makes use of data generated by the Department of Pathology of the University of Hong Kong and Pfizer, Inc.; a full list of the investigators who contributed to the generation of the data is available from ref. 28.

AUTHOR CONTRIBUTIONS

M.J.W. and B.W. contributed to the development of the model. M.J.W. designed and performed computational simulations with support from C.P.B. M.J.W., A.S. and T.A.G. analyzed the data. C.P.B. contributed to the analysis. T.A.G. and A.S. jointly conceived, designed and developed the model, interpreted the results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. Basanta, D. & Anderson, A.R.A. Exploiting ecological principles to better understand cancer progression and treatment. *Interface Focus* **3**, 20130020 (2013).
3. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
4. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
5. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).
6. Polyak, K. Tumor heterogeneity confounds and illuminates: a case for Darwinian tumor evolution. *Nat. Med.* **20**, 344–346 (2014).
7. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
8. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
9. Tabassum, D.P. & Polyak, K. Tumorigenesis: it takes a village. *Nat. Rev. Cancer* **15**, 473–483 (2015).
10. Shou, W., Bergstrom, C.T., Chakraborty, A.K. & Skinner, F.K. Theory, models and biology. *eLife* **4**, e07158 (2015).
11. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).

12. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
13. Jesinghaus, M. *et al.* Distinctive spatiotemporal stability of somatic mutations in metastasized microsatellite-stable colorectal cancer. *Am. J. Surg. Pathol.* **39**, 1140–1147 (2015).
14. Ohta, T. & Gillespie, J.H. Development of neutral and nearly neutral theories. *Theor. Popul. Biol.* **49**, 128–142 (1996).
15. Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421 (1995).
16. Durrett, R. & Schweinsberg, J. Approximating selective sweeps. *Theor. Popul. Biol.* **66**, 129–138 (2004).
17. Driessens, G., Beck, B., Caauwe, A., Simons, B.D. & Blanpain, C. Defining the mode of tumour growth by clonal analysis. *Nature* **488**, 527–530 (2012).
18. Luria, S.E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
19. Griffiths, R.C. & Tavaré, S. The age of a mutation in a general coalescent. *Communications in Statistics* **14**, 273–295 (1998).
20. Maruvka, Y.E., Kessler, D.A. & Shnerb, N.M. The birth-death-mutation process: a new paradigm for fat tailed distributions. *PLoS One* **6**, e26480 (2011).
21. Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.* **23**, 230–250 (2013).
22. Kessler, D.A. & Levine, H. Large population solution of the stochastic Luria-Delbruck evolution model. *Proc. Natl. Acad. Sci. USA* **110**, 11682–11687 (2013).
23. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: an explanation of the 1/*f* noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
24. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA* **105**, 4283–4288 (2008).
25. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA* **107**, 18545–18550 (2010).
26. Sun, S., Klebaner, F. & Tian, T. A new model of time scheme for progression of colorectal cancer. *BMC Syst. Biol.* **8** (suppl. 3), S2 (2014).
27. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
28. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).
29. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
30. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).
31. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
32. Attolini, C.S.-O. *et al.* A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. USA* **107**, 17604–17609 (2010).
33. Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B. & Beerenwinkel, N. The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One* **6**, e27136 (2011).
34. Sprouffske, K., Pepper, J.W. & Maley, C.C. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prev. Res. (Phila.)* **4**, 1135–1144 (2011).
35. Guo, J., Guo, H. & Wang, Z. Inferring the temporal order of cancer gene mutations in individual tumor samples. *PLoS One* **9**, e89244 (2014).
36. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA* **110**, 4009–4014 (2013).
37. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
38. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342**, 995–998 (2013).
39. Heng, H.H.Q. *et al.* Stochastic cancer progression driven by non-clonal chromosome aberrations. *J. Cell. Physiol.* **208**, 461–472 (2006).
40. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
41. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54–58 (2014).
42. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Reports* **6**, 514–527 (2014).

ONLINE METHODS

Data analysis. The processing of exome sequencing data from ref. 11 and TCGA[12] involved variant calling on matched tumor-normal pairs using Mutect[43]. A mutation was considered if the depth of coverage was ≥10 and at least three reads supported the variant. Mutations that aligned to more than one genomic location were discarded. The whole-genome sequencing gastric cancer data[28] were processed using VarScan2 (ref. 44), with the minimum depth of coverage for a mutation being 10× and at least three reads supporting the variant. Non-CRC samples in the TCGA data had mutations called using Mutect according to the pipeline described in ref. 45. MSI in the TCGA colon cancer samples was called using MSIsensor[46]. Annotation was performed with ANNOVAR[47].

To fit the neutral model to allele frequency data, we considered only variants with an allele frequency in the range $[f_{max},f_{min}]$ corresponding to $[t_0,t]$ in equation (2). The low boundary f_{min} reflects the limit for the reliable detectability of low-frequency mutations in next-generation sequencing data, which is on the order of 10% (ref. 43). The high boundary f_{max} is necessary to filter out public mutations that were present in the first transformed cell. In the case of diploid tumors clonal mutations are expected at $f_{max} = 0.5$ (mutations with 50% allelic frequency are heterozygous public or clonal), in the case of triploid tumors this threshold drops to 0.33 and in the case of tetraploid neoplasms it drops to 0.25. For all samples, we used a boundary of [0.12,0.24] to account only for reliably called subclonal mutations and tumor purity in the samples. All the samples considered in this study were reported to have tumor purity ≥70% and a minimum of 12 reliably called private mutations within the fit boundary. Once these conditions were met in a sample, equation (7) was used to perform the fit as illustrated in **Figures 1b** and **2b**. In particular, for $x = 1/f$, equation (7) becomes a linear model with slope μ/β and intercept $-\mu/(\beta f_{max})$. We exploited the intercept constraint to perform a more restrictive fit using the model $y = m(x - 1/f_{max}) + 0$.

Copy number changes (allelic deletion or duplication) can alter the frequency of a variant in a manner that is not described by equation (7). We assessed the impact of copy number alterations (CNAs) on our estimates of the mutation rate in the TCGA colorectal cancer samples by using the paired publically available segmented SNP array data to exclude somatic mutations that fell within regions of CNA. CNAs were identified as having an absolute log R ratio >0.5, and model fitting was performed only on diploid regions of the genome. In the gastric cancer cohort, regions with copy number changes were identified using Sequenza[48] and removed from the analysis. Mutation rates were adjusted to the size of the resulting diploid genome. The robustness of our analysis to copy number changes is demonstrated in **Supplementary Figures 2** and **5**. R^2 values were independent from the mean coverage of mutations ($P = 0.32$), the total number of mutations in the sample ($P = 0.40$), the mutation rate ($P = 0.11$) and the number of mutations within the model range ($P = 0.65$). The sequencing data from our previous publication[11] are accessible via the ArrayExpress database under accession E-MTAB-2247. The TCGA data are accessible via the database of Genotypes and Phenotypes (dbGaP) under accession phs000178.v9.p8. Whole-genome sequencing gastric cancer data are accessible through the European Genome-phenome Archive (EGA) database under accession EGAS00001000597.

Stochastic simulation of tumor growth. To further validate our analytical model and to test robustness to the noise in next-generation sequencing data, we developed a stochastic simulation of tumor growth and accumulation of mutations that allowed us to generate synthetic data sets. The model was written and analyzed in the Julia programming language (http://julialang.org/). We then applied the analytical model to the simulated data to confirm that sources of noise in next-generation sequencing data did not considerably influence our results. In particular, we verified that we could reliably extract the input parameters of the simulation (namely, the mutation rate) from 'noisy' synthetic data. Confounding factors in the data included normal contamination, sampling effects, the detection limit of next-generation sequencing mutation calling and variable read depth. We simulated a tumor using a branching process with discrete generations, beginning with a single 'transformed' cancer cell that gives rise to the malignancy. Under exponential growth, the population at time t will be given by

$$N(t) = R^t = e^{\ln(R)t} \tag{9}$$

where R is the average number of offspring per cell and the time t is in units of generations. We will consider primarily the case when $R = 2$ (a cell always divides into two cells), but we will also consider values <2, noting that R must be greater than 1 to have growth. At each division, cells acquire new mutations at a rate μ, and we assume that every new mutation is unique (infinite sites approximation). The number of mutations acquired by a daughter cell at division is a random number drawn from a Poisson distribution with mean μ. Each cell in the population is defined by its mutations and its ancestral history (by recording its parent cell). Using this information, we can then reconstruct the history of the whole tumor and, crucially, calculate the variant allele frequency of all mutations in the population. To relate the discrete simulation to the continuous analytical model, we will now rederive equation (7) in the context of our computational model. As we simulate a growing tumor using discrete generations, both the mutation rate μ and per capita growth rate $\lambda = \ln(R)$ are in units of generations. For an offspring probability distribution $P = (p_0,p_1,p_2)$ where $p_k = P(\text{number of offspring} = k)$, the average number of offspring R is simply given by the expected value of P

$$R = \mathrm{E}[P] = p_1 + 2p_2 \tag{10}$$

For example, for $R = 2$, we have $P = (p_0 = 0,p_1 = 0,p_2 = 1)$. By choosing different offspring probability distributions, we can easily modulate the growth rate. We note that we are now expressing both μ and λ as rates per generation rather than probabilities (all rates are scaled by units of generation). This allows us to write the growth function as $N(t) = \exp(\lambda t)$ with $\lambda = \ln(R)$. Proceeding as in the main text, our cumulative number of mutations with an allelic frequency f is therefore

$$M(f) = \frac{\mu}{\lambda}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \tag{11}$$

Therefore, when fitting the model to our stochastic simulation, we extract μ/λ from the linear fit, making it straightforward to compare the simulation with the analytical model.

Next-generation sequencing data only capture a small fraction of the variability in a tumor, as the resolution is often limited to alleles with a frequency of >10% because of sequencing depth and limitations in mutation calling. To account for this, we employ a multistage sampling scheme in our simulations. For all simulations reported here, we grow the tumor to a size of 1,024 cells, which gives a minimum allele frequency of ~0.1%, considerably lower than the 10% attainable in next-generation sequencing data. After growing the tumor and calculating the variant allele frequency for all alleles, we take a sample of the alleles in the population, noting that we are assuming that the population is well mixed and has no spatial structure. We can vary the percentage of alleles we sample, thus allowing us to investigate the effect of the depth of sequencing on our results. As we know the true allelic frequency in the simulated population, we can use the multinomial distribution to produce a sample of the 'sequenced' alleles, where the probability of sampling allele i is proportional to its frequency. The probability mass function is given by

$$f(x;n,p) = \frac{n!}{x_1!\ldots x_k!}\prod_{i=1}^{k} p_i^{x_i} \quad , \quad x_1 + \ldots + x_k = n \tag{12}$$

where x_i is the sampled frequency of allele i, n is the number of trials (the chosen percentage of alleles sampled) and p_i is the probability of sampling allele i (which has frequency ρ_i in the original population)

$$p_i = \frac{\rho_i}{\sum_{j=1}^{k}\rho_j} \tag{13}$$

The variant allele frequency (VAF) is therefore given by

$$\mathrm{VAF} = \frac{x_i}{N_i} \tag{14}$$

where N_i is the total number of sampled cells from which every sampled allele is derived. As we are assuming a constant mutation rate μ, we can assume that the percentage of alleles sampled comes from an equivalent percentage of cells. However, to include an additional element of noise that resembles the variability of read depth, we calculate a new N_i for each allele i that

approximates the read depth. For a desired 'sequencing' depth D, we calculate the corresponding percentage of the population we need to sample that will give us our desired depth. For example, for a desired depth of 100× from a population of 1,000 cells, we would need to sample 10% of the population. To include some variability in depth across all alleles, we use Binomial sampling so that N_i is a distribution with mean D.

Contamination from non-tumor cells in next-generation sequencing results in variant allele frequencies being underestimated. To include this effect in our simulation, we can modify our N_i by an additional fraction ε, the percentage of normal contamination. Our variant allele frequency calculation thus becomes

$$\mathrm{VAF} = \frac{x_i}{N_i(1+\varepsilon)}$$

We also include a detection limit in our sampling scheme; we only include alleles that have an allelic frequency greater than a specified limit in the original tumor population.

To include the effects of selection in the simulation, we introduce a second population, where on average each cell has a greater number of offspring than a cell from the first population. To model this, our second population has a modified offspring probability distribution: the previous offspring probability distribution was $P = (p_0,p_1,p_2)$ and the offspring probability distribution of our second, fitter population is defined as $Q = (q_0,q_1,q_2)$, where $q_2 > p_2$. The selective advantage of a population s will be given by the ratio of the expected number of offspring

$$1+s = \frac{\mathrm{E}[Q]}{\mathrm{E}[P]} = \frac{q_1+2q_2}{p_1+2p_2}$$

Therefore, given P and a desired selective advantage s, we can easily calculate the offspring probability distribution of a fitter clone, Q.

Previous studies have detected the presence of mixtures of subclones in breast cancer samples that emerged by means of clonal expansions, thus generating multiple subclonal clusters in the data[31]. We also used our computational model of next-generation sequencing data to produce similar synthetic data by mixing different clonal clusters and verified that, in this scenario (a model of differential selective pressure across subclones), the power law does not hold. The simulation code is available at https://github.com/andreasottoriva/neutral-tumor-evolution.

Simulation results. From the simulated data, we produced histograms of the allelic frequency and calculated $M(f)$ to fit the analytical model. We used the same frequency range as applied to empirical data $[f_{\max},f_{\min}] = [0.12,0.24]$. Equivalent plots to those in **Figure 1a**,**b** but with simulated data are shown in **Supplementary Figure 9a**,**b**. These demonstrate that we are able to accurately model the allelic distribution of next-generation sequencing data with our simple neutral model of tumor growth. We also show the effect of a low mutation rate (**Supplementary Fig. 9c**), a large number of clonal mutations (**Supplementary Fig. 9d**), 30% contamination in the sample (**Supplementary Fig. 9e**) and a low detection limit (**Supplementary Fig. 9f**). Notably, by fitting the analytical model to the simulated data, we can recover the input mutation rate with high accuracy (**Supplementary Fig. 9g**, 10,000 equivalent simulations). The mean percentage error from the fit is 1.1%. We also see uniformly high R^2 values across all simulations (**Supplementary Fig. 9h**).

To test the robustness of the model to the number of clonal mutations, the detection limit and the amount of normal contamination, we ran 10,000 simulations across the spectrum of these parameters (**Supplementary Fig. 10a**,**b**). We accurately recover (to within 15%) the mutation rate for 95% of simulations across different numbers of clonal mutations and different detection limits. In contrast, we found that levels of normal contamination above 30% considerably influence the parameter estimations of the model: hence, our decision of only considering samples with ≥70% tumor content (**Supplementary Fig. 10c**). Indeed, when normal contamination is above 30%, the clonal peak in the allelic frequency distribution interferes significantly with our chosen cumulative sum limit ($f_{\max} = 0.24$), thus affecting our results. Nevertheless, the estimates are within a factor of 2 for normal contamination of up to 50%, which we consider an acceptable level of accuracy. When we consider normal contamination ε directly in our analytical model, the allelic fraction of a new mutation becomes

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda\beta t}(1+\varepsilon)} \tag{15}$$

and, consequently, $M(f)$ is

$$M(f) = \frac{\mu}{\beta(1+\varepsilon)}\left(\frac{1}{f} - \frac{1}{f_{\max}}\right) \tag{16}$$

showing that normal contamination alters the measurement of mutation by a factor of $1/(1+\varepsilon)$: much lower than one order of magnitude. Furthermore, if normal contamination could be estimated accurately from histopathological scoring or from reliable bioinformatics tools, we would be able to correct the frequency of variants in the data and thus rescue our ability to correctly estimate parameters with up to 40–45% normal contamination (**Supplementary Fig. 10d**). We also tested the model with varying read depths and mutation rates. We find that either a low mutation rate or a low read depth results in a higher proportion of poor model fits (R^2 <0.98) and inaccurate or higher variance in mutation estimates (**Supplementary Fig. 10e–h**). It is therefore possible that because of our stringent neutrality criteria the true proportion of tumors that are dominated by neutral dynamics is higher than reported; related to this, our gastric cancer cohort covers the whole genome (greater mutation rate per division) and has mean depth of coverage >90×, which may explain in part why we see a greater proportion of gastric cancers classified as neutral.

Additionally, we tested the model with simulations using a range of different probability distributions for the number of surviving offspring at each cell division. We simulated a growing tumor 10,000 times with five different offspring probability distributions and then reported the distributions of the fitted parameters. As λ decreases, the distribution of mutation estimates becomes wider (**Supplementary Fig. 10i**,**j**), and we see an increase in poorly fitted models (larger number of models with R^2 <0.98). Again, this suggests that tumor growth may still be neutral even when we classify a tumor as non-neutral because of a poor R^2 value. Hence, our underestimation of the number of neutral cases may be largely due to a low proportion of cells that successfully produce two viable offspring (the β term in equation (7)) rather than the presence of selection.

By introducing a second, fitter population early during tumor growth, we show that the fitter clone causes an over-representation of variants at high frequency as compared to what we would expect from our null model of neutral tumor growth. This causes the cumulative distribution to bend and deviate from the linear relationship predicted by neutral growth, as shown in **Supplementary Figure 11a**,**b**. This is because an over-representation of variants at high frequency, as compared to what we would expect from our null model, is caused by clonal selection of the fitter clone, but we note that we do not know what caused this increase (it could be a point mutation, chromosomal aberration or change in environmental pressures, for example). In other words, some passenger mutations are just in the 'right clone at the right time' and become over-represented in the tumor when that 'right' clone expands.

We also show that having multiple subclones that arose by means of clonal expansion, thus producing multiple clonal 'clusters', produces a deviation from the linear relationship we predict (**Supplementary Fig. 11c–f**), as does having a marked increase in the mutation rate early in tumor growth (**Supplementary Fig. 11g**,**h**).

43. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
44. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
45. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* doi:10.1038/nm.3984 (30 November 2015).
46. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
47. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
48. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).