Between-region genetic divergence reflects the mode and tempo of tumor evolution

Ruping Sun1–3,9, Zheng Hu1–3,9, Andrea Sottoriva4, Trevor A Graham5, Arbel Harpak6, Zhicheng Ma1–3, Jared M Fischer7, Darryl Shibata8 & Christina Curtis1–3

Given the implications of tumor dynamics for precision medicine, there is a need to systematically characterize the mode of evolution across diverse solid tumor types. In particular, methods to infer the role of natural selection within established human tumors are lacking. By simulating spatial tumor growth under different evolutionary modes and examining patterns of between-region subclonal genetic divergence from multiregion sequencing (MRS) data, we demonstrate that it is feasible to distinguish tumors driven by strong positive subclonal selection from those evolving neutrally or under weak selection, as the latter fail to dramatically alter subclonal composition. We developed a classifier based on measures of between-region subclonal genetic divergence and projected patient data into model space, finding different modes of evolution both within and between solid tumor types. Our findings have broad implications for how human tumors progress, how they accumulate intratumoral heterogeneity, and ultimately how they may be more effectively treated.

The multistage model of carcinogenesis described in the early 1950s (refs. 1,2) and Nowell’s 1976 perspective piece on the clonal evolution of tumor cells3 provided a conceptual framework for understanding tumor progression. These and other studies4,5 were foundational in defining the elements of somatic evolution. However, the evolutionary dynamics that govern tumor initiation and subsequent growth after transformation remain poorly understood. Moreover, the distinction between stages is often blurred because tumorigenesis is largely occult, often taking place over decades6,7 where lesions are only detected once they achieve a certain size or cause symptoms.

Evolution is the product of three major underlying processes: mutation, selection, and genetic drift8. Mutations are readily measured in human tumors, and it is generally assumed that ongoing strong selection governs the growth of an established tumor after transformation, leaving a detectable signal on the genome, where the acquisition of additional ‘drivers’ results in multiple selective sweeps9,10. In this scenario, driver mutations accompanied by numerous hitchhiking passenger mutations can attain high frequency and manifest as ‘subclonal clusters’ in bulk tumor sequencing data10. This led to the development of a suite of methods aimed at inferring subclonal clusters. However, inference of the number of subclones and their proportions from bulk tumor sequencing is a non-trivial task with the solution non-identifiable under most conditions11–14. Drift can also cause extensive intratumoral heterogeneity (ITH) that may be difficult to distinguish from selection without appropriate population genetics methods. For example, we proposed and tested several predictions of a Big Bang model of colorectal tumor growth, wherein, after transformation, the tumor grows as a single terminal expansion populated by a large number of heterogeneous—and effectively equally fit—subclones15. In this model, most detectable subclonal (private) alterations arise early during growth. While post-transformation selection could be detected in these colorectal tumors, it was often too weak to alter tumor subclonal architecture. Rather, patterns of ITH were suggestive of effectively neutral evolution.

Other studies have since corroborated ‘Big Bang’ dynamics in colorectal tumors16–19. Additionally, neutral evolution was reported in hepatocellular carcinoma via in-depth multiregion profiling20. Williams et al. further investigated evidence for neutral evolution in multiple solid tumors using bulk single-sample sequencing data in comparison with a theoretical null neutral model21. However, as we show, this task is better powered using MRS, which captures additional features of genetic diversity.

Progression modes and tempos differ between neutrally evolving tumors and tumors with post-transformation selection. Hence, there remains a need for the systematic evaluation of different modes of evolution in diverse solid tumors within a population genetics framework. As selection is complex, it is instructive to initially focus on the commonly assumed scenario of strong positive selection after transformation and contrast this with a neutral model. We leverage the fact that spatiotemporal patterns of genetic variation among cancer

1Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. 2Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. 3Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA. 4Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK. 5Evolution and Cancer Laboratory, Barts Cancer Institute, Queen Mary University of London, London, UK. 6Department of Biology, Stanford University, Stanford, California, USA. 7Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, Oregon, USA. 8Department of Pathology, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA. 9These authors contributed equally to this work. Correspondence should be addressed to C.C. (cncurtis@stanford.edu).
cell populations and in particular their variant allele frequency (VAF) distributions (also known as the site frequency spectrum, or SFS)\(^{22}\) derived from next-generation sequencing (NGS) can be used to test hypotheses about the underlying evolutionary processes, including the strength of selection and extent of genetic drift. To this end, we simulated spatial tumor growth under different modes of evolution and trained a classifier on the basis of ITH metrics derived from the SFS to discriminate between these scenarios. By projecting MRS data from various solid tumors into model space, we categorize their patient-specific evolutionary dynamics.

RESULTS
Spatial simulation of distinct modes of tumor evolution
To investigate how different modes of tumor evolution influence the SFS from bulk sequencing data, as well as the power to detect signals of positive selection, we developed an agent-based model of spatial tumor growth (parameters reported in Supplementary Table 1). Within this framework, we simulated various modes of tumor evolution, including a neutral model and an alternate neutral model based on cancer stem cell (CSC)-driven growth (neutral CSC model). We also simulated various levels of positive selection (\(s = 0.01, 0.02, 0.03, 0.05, \) and 0.1), such that the acquisition of advantageous mutations alters the cell birth–death rate according to the selection coefficient \(s\) (Fig. 1, Online Methods, and Supplementary Fig. 1). In all models, random neutral point mutations arise via a Poisson process during each cell division. Virtual tumor growth is simulated via the expansion of deme\(^{23}\) subpopulations (neighborhoods of 5,000–10,000 cells) within a defined 3D lattice, and cells within each deme are well mixed and replicate via a random branching process. By recording mutational lineages as the tumor expands and subsequently virtually sampling the ‘final’ tumor, as is done experimentally after resection, we evaluate differences in the SFS arising under different levels of selection and the utility of different tissue sampling strategies (Fig. 1a). Thus, we model spatial tumor growth and the inherent stochasticity of this process while accounting for the truncated SFS derived from bulk sequencing due to the large number of rare subclones that are not sampled or are below detection limits. This facilitates comparisons with data derived from patient tumors analyzed within a sensitive pipeline for calling somatic single-nucleotide variants (SSNVs) from MRS (Fig. 1b, Online Methods, and Supplementary Fig. 2). A summary of terminology is provided in Supplementary Table 2. A summary of the MRS data sets included in this study is provided in Supplementary Table 3.

Spatial subclone composition and the distribution of subclonal VAFs derived from MRS (\(n = 2, 4, \) and 8 regions) of virtual tumors differed dramatically depending on the mode of evolution, as illustrated for representative virtual tumors (Fig. 2a,b and Supplementary Fig. 3). In particular, under stronger selection (\(s \geq 0.02\)), multiple subclone expansions occur in different regions of the virtual tumor, as shown in the clone map (Fig. 2a). Likewise, multiple peaks (mutational clusters) were observed in the SFS histograms due to the enrichment of high-frequency (VAF > 0.2) subclonal SSNVs under stronger selection (Fig. 2b and shown schematically in Supplementary Fig. 4), which were largely region specific, reflecting elevated genetic divergence. Indeed, subclonal selection typically resulted in detectable differences in the SFS histograms from different tumor regions.

Figure 1 Overview of the simulation framework and the genomic data analysis pipeline. (a) Schematic overview of our agent-based computational framework to simulate 3D tumor growth (after transformation) under various modes of evolution, including neutral evolution and different levels of positive selection, followed by spatial sampling and MRS of the virtual tumor. Tumor growth is simulated via the expansion of deme subpopulations within a defined 3D cubic lattice according to explicit rules dictated by spatial constraints, where cells within each deme are well mixed and grow via a stochastic branching (birth–death) process (Online Methods and Supplementary Fig. 1). By simulating the acquisition of random mutations (neutral or beneficial), tracing the genealogy of each cell as the tumor expands, and subsequently virtually sampling and sequencing the final virtual tumor, as is done experimentally after resection or biopsy, it is possible to evaluate differences in the SFS under different modes of selection and sampling strategies. Five ITH metrics derived from the SFS were employed to distinguish between different evolutionary modes. Sub muts, subclonal mutations. (b) A unified sequencing analysis pipeline based on SSNV calling and copy number estimation, as well as stringent quality control, was employed to obtain VAF estimates adjusted for purity and local copy number for seven MRS data sets derived from patient samples across diverse tissue types. The ITH metrics were similarly computed in patient tumor samples and compared to those observed in virtual tumors under different evolutionary modes. Pvt, private.
In contrast, under neutral growth, a neutral CSC-like model where only a subset of cells have unlimited proliferative potential (equivalent to a smaller deme size), or weak selection (s = 0.01), subclonal composition is maintained in the final tumor. The SFS for these three modes were generally similar between regions consisting of two mutational clusters, namely a public cluster centered at VAF = 0.5 composed of mutations that occurred before transformation and present in all tumor cells (fixed) and a right-skewed distribution of private subclonal mutations at low VAF (<0.25) (Fig. 2b), where their detection depends on sequencing depth. Notably, MRS but not single-sample sequencing enables the identification of private SSNVs present at high frequency in one or a few regions but subclonal in the entire tumor (Supplementary Fig. 4). Indeed, at least two spatially separated regions are needed to accurately distinguish public SSNVs in solid tumors, as mutations that are subclonal in the whole tumor can appear clonal within some samples owing to sampling bias24. In each of the modes, over 70% of subclonal SSNVs were region specific as a result of spatial constraints during virtual tumor expansion. However, selection increased the fraction of high-frequency (VAF > 0.2) region-specific subclonal SSNVs out of all region-specific subclonal SSNVs (VAF > 0.08) (fHs) (Supplementary Table 4). Hence, MRS aids the identification of subclonal SSNVs that reflect the dynamics of clonal expansion after tumor transformation, whereas clonal SSNVs are not informative in this regard.

To quantify the extent of ITH defined as between-region genetic divergence on the basis of subclonal SSNVs (identified through MRS) under different levels of selection, we employed the following metrics (Online Methods) in addition to fHs (defined above):

- fHs: the fraction of subclonal SSNVs (VAF > 0.08) with high frequency (VAF > 0.2);
- FST (fixation index): a measure of genetic divergence between regions;
- KSD (Kolmogorov–Smirnov distance): dissimilarity of the SFS between regions.

As expected, the fHs and fHs were correlated, as were other features, albeit to a lesser extent (Supplementary Fig. 5). All of the statistics increased in value under stronger selection (s ≥ 0.02) relative to the neutral, neutral (CSC), and weak selection (s = 0.01) models. This suggests that selection causes characteristic and detectable genetic divergence between regions when it fails to result in complete sweeps (Fig. 2b and Supplementary Table 4).

We further explored the relationship between different modes of evolution and genetic divergence captured by MRS (n = 2, 4, or 8 regions) and single-sample sequencing at various depths (80–640×)
(Fig. 2c, Online Methods, and Supplementary Figs. 6 and 7). For reference, the theoretical cumulative SFS assuming neutral exponential growth in a well-mixed population21,26 (referred to as the theoretical neutral SFS) is also shown. Differences in the SFS were evident such that tumors simulated under higher selection (s ≥ 0.02) typically fell above the theoretical neutral SFS, whereas the remaining modes generally traced or fell below this curve. The variability in the SFS within individual modes highlights the importance of stochastic simulations.

To compare the utility of single-sample data versus MRS, we computed the ratio of the area under the cumulative SFS (based on the pooled VAFs for MRS) to the area under the theoretical neutral SFS (rAUC), as this is applicable to both single-sample sequencing and MRS. Comparison of the rAUC for virtual tumors simulated under different modes demonstrates the challenge of distinguishing between s > 0.05 and s ≤ 0.01 (including the neutral and neutral (CSC) models) with a single sample, even at high depth, whereas better separation is achieved with even one additional region (Supplementary Fig. 8). This is also reflected in comparisons of sensitivity and specificity to distinguish alternative models from the simulated neutral model on the basis of the rAUC (Supplementary Fig. 9a). Whereas power increased with selection intensity (s = 0.05–0.1) and the number of regions (n = 2–8), this was not the case for increased depth alone, owing to sampling bias and the inability to capture regionally localized high-frequency subclonal mutations that arose under strong selection (Online Methods and Supplementary Fig. 9b). In contrast, metrics that capture between-region ITH such as fHsub are better able to distinguish a specific alternate model than rAUC. Of note, s = 0.01 could not be distinguished from the simulated neutral model. The neutral (CSC) model is also similar to the ‘vanilla’ neutral model but generates localized diversity. Thus, we refer to these three modes as effectively neutral, as the population dynamics of such nearly neutral mutations are virtually equivalent to those of neutral mutations27,28. Similarly, it was not feasible to distinguish the SFS under different levels of elevated selection (s ≥ 0.02) (Supplementary Fig. 5). Many factors can dampen signals of selection, as in the case of strong but less frequent drivers that are very rare or occur late without sufficient time to expand (Supplementary Fig. 10). As such, we focus on effective neutrality and strong selection (s ≥ 0.02) but present results from all modes for completeness.

The site frequency spectrum reflects tumor growth dynamics

To evaluate the SFS in patient samples, we first analyzed MRS data from colorectal adenocarcinomas sampled from two regions (COAD, taken ≥ 3 cm apart)15 with high purity (72–96%) and adequate coverage (80–120× median whole-exome sequencing (WES) depth) (Supplementary Figs. 11 and 12). We devised a MuTect-based Variant Assurance Pipeline (VAP) to enable the sensitive and accurate detection of subclonal SSNVs from MRS (Online Methods, Supplementary Figs. 2 and 13, and Supplementary Note). The observed VAF estimates were adjusted for sample purity and local copy number, enabling pairwise comparisons between tumor regions, and throughout we refer to adjusted VAFs as VAFs (Supplementary Figs. 14 and 15). As noted above, the SFS histograms appear bimodal for both regions, as shown for representative tumors spanning the major pathways of colorectal cancer pathogenesis, categorized according to microsatellite instability (MSI) versus microsatellite stability (MSS) status and chromosomal instability (CIN) status29 (Fig. 3a). A peak centered at a VAF of 0.5 was observed in all tumors with constituent mutations that were present at similar frequencies in the left and right samples (Fig. 3b). This VAF cluster primarily represents public mutations present in the founding tumor cell. Whereas private high-VAF (0.2–0.4) SSNVs were infrequent, low-frequency subclonal SSNVs (VAF < 0.2) were common and were generally region specific despite having similar VAFs, suggesting that mutation frequency is not a reliable surrogate for subclone identity. Similar patterns were observed in additional cancers and an adenoma (Supplementary Fig. 15). We computed the five ITH metrics, which exhibited low or intermediate values for tumors M, O, and U comparable to those noted in virtual tumors under effectively neutral growth. In contrast, tumors G, N, and W and adenoma S exhibited higher values, similar to those noted in virtual tumors subjected to selection (Fig. 3 and Supplementary Tables 4 and 5).

We further evaluated the genetic divergence within a clonal in vivo tumor growth model by generating single-cell expansions from mismatch repair (MMR)-deficient COAD cell lines followed by xenotransplantation into the opposite flanks of immune-compromised mice and WES of the resultant tumors (Online Methods). In both technical replicates and independent cell line experiments, the data yielded SFS histograms that lacked enrichment for high-frequency private SSNV’s (Supplementary Figs. 16 and 17). Additionally, the corresponding ITH metrics were congruent with effectively neutral growth, as might be expected for fully transformed cells that do not require further alterations to propagate tumor growth.

VAF clusters do not necessarily capture subclone identity

Existing computational methods to infer tumor subclonal architecture from bulk sequencing data exploit the observation that SSNVs cluster around several distinct VAF modes or clonal clusters10–13,30. These methods aim to assign subclone identity on the basis of the assumption that mutations with similar frequencies are in the same cell and that a limited number of dominant subclones underwent clonal expansion9,11,31. However, mutational clusters do not guarantee unique lineages and therefore do not necessarily capture clonal identity. In addition, subclone architecture is influenced by selection and spatial constraints. Indeed, visual inspection of the SFS histograms and scatterplots from the bi-sampled COAD data set showed that, in all cases, the majority of subclonal SSNVs with VAF < 0.2 region specific (Fig. 3 and Supplementary Fig. 15). This suggests that mutations grouped on the basis of their VAFs do not correspond to unique clones. To evaluate subclonal architecture at higher resolution, we performed WES on five individual COAD glands and bulk samples from two distant tumor regions of a representative cancer (COAD-O). The private mutations specific to either bulk sample (OA or OB; Fig. 4a,b) were only detected in glands from the same tumor region (P = 5 × 10−11, Fisher’s exact test), and similar patterns were noted based on targeted sequencing of private SSNVs in multiple individual glands for each of the bi-sampled COADs (Supplementary Fig. 18). In a subset of single glands from two spatially separated regions, the same SSNVs were detected despite being subclonal in the bulk tumor (Fig. 4b, green dots), potentially reflecting early subclone mixing15,19 or sampling of a clone boundary. In contrast, later-arising SSNVs were generally region specific, consistent with spatial constraints during expansion. SSNVs specific to bulk sample OA (VAF < 0.2) were detected in different combinations of single glands with VAF > 0.2, suggesting that distinct lineages can have similar VAFs in the bulk tumor. Reconstruction of a possible phylogenetic tree using LiChE32 also demonstrated subclone spatial segregation, where essentially every gland within a bulk region is a subclone (Fig. 4d), emphasizing the star-like phylogeny predicted for a neutrally growing population26 (Supplementary Fig. 19). WES of single glands from COAD-U yielded similar results (Supplementary Fig. 20).
We further reasoned that a ‘true’ clone should form a cluster that persists (for example, where mutations remain grouped), irrespective of the inclusion of data from additional regions. We evaluated this in other solid tumors by analyzing published MRS data sets for esophageal carcinoma (ESCA)\(^{33}\), lung adenocarcinoma (LUAD)\(^{34}\), non-small-cell lung cancer (NSCLC)\(^{35}\), glioma (GLM)\(^{36}\), and glioblastoma (GBM)\(^{37}\) (Online Methods, Supplementary Figs. 2 and 11, and Supplementary Table 3). Application of SciClone\(^{13}\) to MRS data from several representative tumors (COAD-O, ESCA-8, and LUAD-4990 for which 2, 3, and 4 regions were available, respectively) consistently resulted in the dissolution of subclonal clusters when data from additional regions were included in the analysis (Fig. 4c and Supplementary Figs. 21–23). Whereas SSNVs in the subclonal clusters did not remain grouped, those in the clonal clusters did (\(P = 0.0003\), Fisher’s exact test), consistent with these mutations being in the founding clone. A persistent mutational cluster in LUAD-4990 was detected through the analysis of four regions, potentially corresponding to a subclone that arose under selection (Supplementary Fig. 22). Collectively, these results illustrate conceptual challenges in inferring subclonal architecture from bulk sequencing VAF data alone.

**Distinguishing the mode and tempo of solid tumor evolution**

We next evaluated genetic divergence based on MRS for treatment-naive primary tumors, including COAD, ESCA, LUAD, LUSC, and GBM, relative to that observed in virtual tumors under different modes. Non-hypermutated GBMs (\(n = 2\)) and GLMs (\(n = 2\)) obtained before and after treatment with temozolomide, a mutagenic alkylating agent...
assumed to impose a positive selective pressure\textsuperscript{36}, were included as positive controls. Additionally, matched Barrett’s esophagus (BE) lesions and adenocarcinomas from two patients (BE-ESCA-4 and BE-ESCA-14) were included as positive controls, as selection is expected during progression from a premalignant lesion. The degree of deviation of the pooled cumulative SFS above the theoretical neutral curve highlights differences in selection across tumor types (Fig. 5a). As predicted, each of the positive controls exhibited cumulative SFS above the neutral curve, consistent with strong selection. In contrast, deviation below the theoretical neutral curve is indicative of spatial constraints, as illustrated by simulating smaller deme sizes (500–1,000 versus 5,000–10,000 cells), where the ability to distinguish selection from effective neutrality was reduced (Supplementary Figs. 24 and 25). Such strong spatial constraints result in infrequent sharing of subclonal mutations between regions (f Shr; Supplementary Tables 4 and 5), a pattern inconsistent with most patient tumors (\( P < 2.2 \times 10^{-16} \), Wilcoxon rank-sum test), suggesting that larger deme size better reflects the patient data.

COAD-M and ESCA-14 exhibited bimodal SFS histograms with scant enrichment for high-frequency private SSNVs, most consistent with patterns of effective neutrality (Fig. 5b). In contrast, COAD-N and LUAD-270 exhibited modest enrichment for such SSNVs, whereas this was more striking in ESCA-8 and LUAD-4990 (Fig. 5c). Despite the lower number of SSNVs in treatment-naive primary GBMs, enrichment of high-frequency private SSNVs was evident and similar to that noted in the primary versus post-treatment recurrence (Fig. 5b).

The five ITH metrics were calculated for primary solid tumors, paired GLMs and GBMs from before and after temozolamide treatment (positive controls), and BE-ESCA pairs (positive controls), as well as for virtual tumors simulated under various evolutionary modes (Fig. 6a). Among the virtual tumors, all five metrics increased markedly under selection (\( s \geq 0.02 \)) relative to effective neutrality. The primary COADs and ESCAs tended to exhibit lower detectable divergence than lung and brain cancers, which exhibited lower divergence than the temozolamide-treated positive controls.

---

**Figure 4** Single-gland WES shows spatial constraints among subclonal mutations. (a) Pairwise histogram of the SFS and SSNV scatterplots from two regions of COAD-O (OA versus OB). (b) Intersection of SSNVs found in bulk regions and single glands. In the inset, the VAFs for SSNVs specific to single glands versus bulk sample OA (side A) are shown. OA-specific SSNVs present in different sets of single glands collapse to similar VAF values (<0.2) in the bulk sample (blue lines connecting the inset), indicating that mutational clusters do not necessarily guarantee clonal identity. (c) The pooled VAF (derived from four regions) is shown for LUAD-4990, indicating a clonal cluster (centered at 0.5) and two subclonal clusters. In pairwise comparisons of the VAF from two regions (P3 and P1), the clonal VAF cluster persists, consistent with the mutations in this cluster being present in all cells, whereas the subclonal clusters partition into distinct clusters according to the two tumor regions. (d) Phylogenetic tree based on SSNV presence/absence in single glands and bulk samples constructed using LICHeE. The bulk sample and corresponding single glands from the same tumor region have a common lineage relationship, potentially reflecting spatial constraints during tumor expansion. SSNVs in known and candidate driver genes are labeled. A truncal APC indel was also detected but was not used for tree construction.
The SFS is commonly used in population genetics, and it is appreciated that tests of neutrality based on a single summary statistic can be difficult to establish, whereas composite metrics can aid the detection of selection. Given the multifaceted nature of ITH and the noise in real data, we reasoned that the major components of the ITH metrics would capture complementary aspects of subclonal genetic divergence. Independent component analysis (ICA) using the five ITH metrics identified two distinct clusters, corresponding to...

**Figure 5** The SFS reflects differential modes of evolution within and between tumor types. (a) Cumulative SFS based on the merged VAF for tumors derived from four tissue types (colon, esophagus, lung, brain) analyzed using the VAP (Online Methods). For all samples, WES data were utilized, with the exception of the ESCA and BE cases, for which WGS data were available. Each line corresponds to a Bezier-smoothed curve of the cumulative SFS. Thick gray curves correspond to the theoretical cumulative SFS under neutral exponential growth in a well-mixed population, shown for reference. Dashed lines correspond to comparisons of tumor regions sampled at distinct stages of tumor progression in the same patient, for example, BE versus ESCA or treatment-naive primary tumor versus post-treatment (Tx) recurrent brain tumors, both of which represent positive controls for selection. (b) Pairwise SFS histograms from representative tumors of different tissue type are shown and depict the number of SSNVs detected at a given VAF for two regions, where SSNVs are grouped according to public (gray), private, shared (green), and private, region-specific (blue) mutations (as in Fig. 3). Histogram bin widths were optimized on the basis of the number of SSNVs (Online Methods). (c) Two-way density plots of the SSNVs present in each region at a given VAF are shown for two tumors. Non-silent SSNVs in known and candidate driver genes are labeled. The color scale reflects the relative density of mutations.
Figure 6. Projection of patient samples onto distinct evolutionary modes. (a) Violin plots for each of five ITH metrics, namely $f_{\text{Hsub}}, f_{\text{Hrs}}, \text{FST}, \text{KSD},$ and $r\text{AUC}$. Colored violin plots show the virtual tumors simulated under different evolutionary modes, whereas the white plots correspond to patient tumor data. Paired pretreatment primary and post-treatment recurrent brain tumors are denoted by “Tx” and serve as a positive control for selection. (b) ICA of virtual and patient tumors based on the five ITH metrics. The ICs separate virtual tumors simulated under effectively neutral growth (neutral, neutral (CSC), and $s = 0.01$) versus positive selection ($s \geq 0.02$), where the decision boundary for an SVM trained on two ICs based on the virtual tumors (effectively neutral versus positive selection models) is indicated by the dashed line. Large transparent colored circles represent values from virtual tumors under different models (200 tumors from each of the seven modes are shown). Small circles represent patient tumors labeled by their corresponding sample ID and colored according to the type of sample. COAD, colorectal adenocarcinoma; CRA, colorectal adenoma; ESCA, esophageal adenocarcinoma; BE, Barrett’s esophagus; LUAD, lung adenocarcinoma; NSCLC, non-small-cell lung cancer; GLM, glioma; GBM, glioblastoma; Xeno, COAD cell line xenograft. (c) The ratio of private SSNVs at more functional (MF) relative to less functional (LF) sites (dMF/dLF) based on PolyPhen-2 was calculated for each of the primary tumors to evaluate the correlation with various ITH metrics.

We then classified patient tumors and visualized them in model space (Fig. 6b, Supplementary Figs. 28 and 29, and Supplementary Table 5), identifying trends with respect to the mode of evolution in a given tumor type despite patient-to-patient variability. For example, COADs exhibited both effective neutrality and selection, as did ESCAs. In contrast, lung and brain tumors tended to show stronger signals of selection. In total, 5 primary tumors were categorized as being compatible with effective neutrality and 12 were categorized as being compatible with selection, whereas only 3 did not robustly fit either scenario. As expected, all four GBMs and GLMs from before and after temazolamide treatment were most compatible with strong positive selection and several appear as outliers on the ICA, potentially because the full impact of treatment is not modeled (Fig. 6b). The paired BE-ESCA cases (ESCA-BE-14 and ESCA-BE-4) exhibited patterns consistent with selection during tumorogenesis followed by effectively neutral growth of the primary tumor (ESCA-14 and ESCA-4). Patterns of genetic divergence in multiple BE lesions from patient 4 (BE-4) were similarly indicative of selection (Supplementary Fig. 30). Notably, irrespective of whether whole-genome sequencing (WGS) or WES data were used, the classification was the same, indicating that WES is adequate for this task given sufficient subclonal SSNVs (Fig. 6b).

Positive selection for drivers during tumor expansion is expected to be associated with an increase in the rate of private SSNVs at more functional (MF) relative to less functional (LF) sites\(^8\). Among primary
tumors, the dMF/dLF ratio was positively correlated with several ITH metrics, for example, fHsub, FST, and tAUC (Fig. 6c). This suggests a general trend between selection and the levels of detectable between-region genetic divergence, although specific patterns could be model dependent (Supplementary Fig. 31). Conversely, the fold enrichment for driver genes among non-silent public SSNVs was negatively correlated with fHsub, consistent with a greater number of public drivers in tumors characterized by effectively neutral growth (Supplementary Fig. 32). Hence, these results corroborate our finding that patterns of genetic divergence in MRS inform the mode and drivers of tumor growth.

DISCUSSION

Here we show that tumors evolving nearly neutrally or through strong selection exhibit fundamentally different patterns of ITH and that these can be distinguished via MRS. Further, we developed a classification framework based on features of the SFS that capture between-region subclonal divergence and applied this to publicly available MRS data, identifying different modes of evolution within and between solid tumor types. We note that compatibility with effective neutrality does not necessarily imply the complete absence of selection. Rather, positive selection may have been weak, variable, or abrogated by negative selection throughout tumor growth43, but the overall patterns do not deviate significantly from those expected under a neutral model. The timing of a mutation is also critical as, within a rapidly expanding adaptive population, only mutations that occur early are likely to be ‘fixed’ in relevant time frames and detectable by NGS, even if they are under strong positive selection, whereas partial sweeps are potentially common42. The lack of evidence for ongoing stringent selection in some of the tumors examined here is congruent with a Big Bang model of effectively neutral tumor growth where the tumor grows as a single expansion with selection uniformly conferred by common drivers in the first tumor cell15.

The finding that human tumors can be categorized into different modes of evolution has implications for defining the drivers of growth and treatment strategies. For example, nearly neutrally evolving tumors show enrichment for drivers among public SSNVs, and it is potentially most efficacious to target these truncal mutations. While most detectable ITH occurs early during effectively neutral growth, the large number of heterogeneous subclones that fall below detection limits increases the chance that pre-existing treatment-resistant variants are present. In contrast, putatively functional private variants were enriched among tumors characterized by ongoing positive selection, suggesting that these may represent relevant targets.

Our findings also inform practical guidelines for studies of tumor evolution. For example, we show that, while at least two regions are required to robustly distinguish public versus private alterations, inclusion of sequencing data from additional regions yielded greater discrimination between different modes of evolution and was more informative than deeper sequencing of a single sample. Even under strong spatial constraints such as small (500–1,000 cells) deme size, where the efficacy of selection is impeded, sequencing additional regions should aid the detection of selection. Improved sensitivity to distinguish different modes of evolution may be achieved by modeling the distinct architecture and microenvironments in different tissues, although these are as of yet poorly understood43. It will also be important to understand the contribution of deleterious passenger alterations44 and clonal cooperation45,46 to tumor dynamics, as well as to evaluate more complex modes of selection in human tumors. Thus, although MRS does not fully resolve the SFS, it nonetheless captures global and local genetic divergence, enabling the detection of signals of selection in individual tumors under certain conditions.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

This work was funded by awards from the NIH (R01CA182514), the Susan G. Komen Foundation (IRR13260750), and the Breast Cancer Research Foundation (BCRF-16-032) to C.C. and an award from the NIH (R01CA185016) to D.S. Z.H. is supported by an Innovative Genomics Initiative (IGI) Postdoctoral Fellowship. S. is supported by the Chris Rokos Fellowship. T.A.G. was supported by Cancer Research UK. This work was supported in part by NIH P30 CA124435 using the Genetics Bioinformatics Service Center within the Stanford Cancer Institute Shared Resource. The results are in part based upon data generated from the following studies: EGAD00001001394, EGAD00001000714, EGAD00001009900, EGAD00001009984, and EGAD00001001113. We thank members of the Curtis laboratory for helpful discussions.

AUTHOR CONTRIBUTIONS

R.S., Z.H., and C.C. designed the study. R.S. analyzed and visualized the data and performed statistical analyses. Z.H. performed simulation studies. Z.M. and D.S. generated data. R.S., Z.H., and C.C. interpreted the data. A.S. and T.A.G. contributed to earlier analysis of the COAD data set. A.H. provided statistical advice. J.M.F. performed xenograft experiments. D.S. and C.C. provided reagents and data. C.C. supervised the study and wrote the manuscript with input from R.S. and Z.H. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Nordling, C.O. A new theory on cancer-inducing mechanism. Br. J. Cancer 7, 68–72 (1953).
2. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br. J. Cancer 8, 1–12 (1954).
3. Nowell, P.C. The clonal evolution of tumor cell populations. Science 194, 23–28 (1976).
4. Cairns, J. Mutation selection and the natural history of cancer. Nature 255, 197–200 (1975).
5. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–767 (1990).
6. Tsao, J.L. et al. Genetic reconstruction of individual colorectal tumor histories. Proc. Natl. Acad. Sci. USA 97, 1236–1241 (2000).
7. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc. Natl. Acad. Sci. USA 110, 1999–2004 (2013).
8. Hu, Z., Sun, R. & Curtis, C. A population genetics perspective on the determinants of intra-tumor heterogeneity. Biochim. Biophys. Acta http://dx.doi.org/10.1016/j.bbadis.2017.03.001 (2017).
9. Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546–1558 (2013).
10. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).
11. Fischer, A., Vázquez-García, I., Illingworth, C.J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. Cell Reports 7, 1740–1752 (2014).
12. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398 (2014).
13. Miller, C.A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLOS Comput. Biol. 10, e1003665 (2014).
14. Deshwar, A.G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 16, 35 (2015).
15. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. Nat. Genet. 47, 209–216 (2015).
16. Uchi, R. et al. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. PLoS Genet. 12, e1005778 (2016).
17. Sievers, C.K. et al. Subclonal diversity arises early even in small colorectal tumors and contributes to differential growth fates. Gut http://dx.doi.org/10.1136/gutjnl-2016-312232 (2016).
18. Bozic, I., Gerold, J.M. & Nowak, M.A. Quantifying clonal and subclonal passenger mutations in cancer evolution. PLOS Comput. Biol. 12, e1004731 (2016).
19. Suzuki, Y. et al. Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer. Mol. Oncol. 11, 124–139 (2017).
20. Ling, S. et al. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc. Natl. Acad. Sci. USA 112, E6496–E6505 (2015).

21. 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).

22. Bustamante, C.D., Wakeley, J., Sawyer, S. & Hartl, D.L. Directional selection and the site-frequency spectrum. Genetics 159, 1779–1788 (2001).

23. Ray, N., Currat, M. & Excoffier, L. Intra-deme molecular diversity in spatially expanding populations. Mol. Biol. Evol. 20, 76–86 (2003).

24. Siegmund, K. & Shibata, D. At least two well-spaced samples are needed to genotype a solid tumor. BMC Cancer 16, 250 (2016).

25. Holsinger, K.E. & Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting $F_{ST}$. Nat. Rev. Genet. 10, 639–650 (2009).

26. Kimura, M. Model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA 76, 3440–3444 (1979).

27. Grossman, S.R. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327, 883–886 (2010).

28. Ostrow, S.L., Bashir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. eLife 3, e10042 (2014).

29. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The ecology and evolution of cancer: the ultra-microevolutionary process. Annu. Rev. Genet. 50, 347–369 (2016).

30. Qiao, Y. et al. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. Genome Biol. 15, 443 (2014).

31. Li, B. & Li, J.Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. Genome Biol. 15, 473 (2014).

32. Popic, V. et al. Fast and scalable inference of multi-sample cancer lineages. Genome Biol. 16, 91 (2015).

33. Ross-Innes, C.S. et al. Whole-genome sequencing provides new insights into the clonal architecture of Barrett’s esophagus and esophageal adenocarcinoma. Nat. Genet. 47, 1038–1046 (2015).

34. Zhang, J. et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science 346, 256–259 (2014).

35. de Bruin, E.C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science 346, 251–256 (2014).

36. Johnson, B.E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science 343, 189–193 (2014).

37. Kim, H. et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. Genome Res. 25, 316–327 (2015).

38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595 (1989).

39. Grossman, S.R. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327, 883–886 (2010).

40. Ostrow, S.L., Bashir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. eLife 3, e10042 (2014).

41. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The ecology and evolution of cancer: the ultra-microevolutionary process. Annu. Rev. Genet. 50, 347–369 (2016).

42. Messer, P.W. & Petrov, D.A. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28, 659–669 (2013).

43. Lloyd, M.C. et al. Darwinian dynamics of intratumor heterogeneity: not solely random mutations but also variable environmental selection forces. Cancer Res. 76, 3136–3144 (2016).

44. McFarland, C.D., Korolev, K.S., Kryukov, G.V., Sunyaev, S.R. & Mirny, L.A. Impact of deleterious passenger mutations on cancer progression. Proc. Natl. Acad. Sci. USA 110, 2910–2915 (2013).

45. Marusyk, A. et al. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514, 54–58 (2014).

46. Cleary, A.S., Leonard, T.L., Gestl, S.A. & Gunther, E.J. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. Nature 508, 113–117 (2014).
ONLINE METHODS

Multiregion sequencing studies. We evaluated patterns of ITH in several publicly available MRS data sets spanning multiple tumor types, including colorectal adenoma/COAD (one adenoma, six patients with COAD)31, ESCA/BE (three patients)32, LUAD (four patients)33, NSCLC (one patient)33, GLM (three patients)34, and GBM (two patients)35. Numbers refer to cases with MRS data that passed quality control. The study accession IDs and list of samples that met coverage and purity requirements are reported in Supplementary Table 3. Details on sequencing depth and purity are provided in Supplementary Figure 11. All samples were analyzed using a custom pipeline (Supplementary Note) to enable the sensitive detection of private SSNVs and standardized comparisons across cohorts, as detailed below.

Single-gland whole-exome sequencing. Building on our previous description of multiregion WES of colorectal tumors and targeted single-gland sequencing, we performed WES of multiple single glands from two tumors in this study (Fig. 4 and Supplementary Fig. 20) on the Illumina platform using the Agilent SureSelect 2.0 or Illumina NRCEx kit. Samples were collected under an institutional review board (IRB)-approved protocol (University of Southern California Keck School of Medicine) as deidentified excess tissues not requiring patient-specific consent, as previously described36. The single-gland WES data were analyzed using the same pipeline as was applied to bulk tumor regions. Intersection plots for SSNVs found in bulk regions and single glands were generated on the basis of mutations that (i) were covered by at least 20 reads in each sample; (ii) had a VAF above 1.5% in the bulk sample or above 15% in the single glands; and (iii) did not derive from regions with varying patterns of loss of heterozygosity (LOH) among samples.

In vivo modeling of colorectal tumor growth. Cells were expanded in vitro, and a single ‘founding’ cell from this population was cloned and expanded to ~6 million cells before transplantation of ~1 million cells into the right and left flanks of an NSG mouse (HCT116 cells) or a nude (Nu/Nu) mouse (LoVo cells), where tumors were allowed to develop to a size of ~1 billion cells (1 cm^3) before being sampled and subjected to WES (Fig. 3 and Supplementary Figs. 16 and 17). The HCT116 and LoVo mismatch repair (MMR)-deficient COAD cell lines were obtained from the ATCC (authenticated using cytochrome c oxidase I assays and STR typing and tested for mycoplasma contamination) and cultured under standard conditions. Tissue was collected separately from the right and left tumors, and DNA was extracted for WES using the Illumina TruSeq Exome kit, as was DNA from the first-passage population (a polyclonal cell line) for each sample. Procedures performed on the mice were approved by the Institutional Animal Care and Use Committee (IACUC) at the Oregon Health and Science University (OHSU; NSG mice) and the University of Southern California (USC; nude mice).

Somatic SNV calling. SCNA detection, and VAF adjustment. To facilitate quantitative comparisons of the SFS, we devised a unified variant assurance (filtering and rescuing) pipeline (VAP) to achieve balance in sensitivity and specificity when MRS data are available such that information can be borrowed across tumor regions. For each raw VAP call from MuTect (v1.1.4, filtered)37, the read alignment features from all samples were reindexed in an automated fashion to assess the confidence (in detected samples) and evidence (in undetected samples) for the alternative allele (Supplementary Fig. 13). Somatic copy number alterations (SCNAs) and tumor purity (p) were estimated with TitanCNA38 (version 1.8.0) in exome-seq mode (except for the ESCA data set, where WGS data were available). The observed VAF for each detected SSNV was adjusted on the basis of cancer cell fraction (CCF) for the ESCA data set, where WGS data were available. For each detected SSNV, the CCF was estimated as the ratio of the SSNV VAF to the total VAF for all detected SNVs, as calculated by the TitanCNA tool. The CCF was then used to adjust the VAF for each detected SSNV, resulting in a more accurate estimation of the tumor purity.

Spatial computational modeling of tumor growth dynamics. We extended our previously described spatial agent-based model15 to simulate tumor growth and mutation accumulation under different scenarios ranging from neutral evolution to strong selection and compare the SFS of SSNVs arising from one, two, four, and eight regions sampled from the spatially separated quadrants of individual virtual tumors. In this agent-based model, spatial tumor growth is simulated via the expansion of deme subpopulations (composed of 5,000–10,000 cells), which mimics the glandular structures often found in epithelial tumors (Supplementary Table 1). The deme model is well established for modeling spatially expanding populations23. Here deme subpopulations expand within a defined 3D cubic lattice (Moore neighborhood, 26 neighbors), where demes expand by particular rules of spatial constraint (peripheral growth49 or alternatively shifting growth15) while cells within each deme are well mixed and grow via a random branching (birth–death) process. The panmixia of cells in the formation of the first deme from a single transformed cell allows for subclone mixing among early-arriving mutations34,19, which can subsequently spread during tumor expansion. Random neutral mutations arise via a Poisson process at each cell division, assuming an infinite sites model.

More specifically, at each time step, we simulate deme division by selecting a deme at random and choosing a neighboring lattice site where the new deme will be placed. We employ a peripheral growth model34 (Supplementary Table 1) where only demes on the surface of the tumor can grow and divide, such that a random empty neighbor site was chosen for each newly formed deme. The peripheral growth model is supported by recent studies indicating that cancer cells at the periphery of the tumor exhibit higher proliferative activity than those at the core43. We assume a maximum deme size of 10,000 cells to minimize the effect of deme structure, which hinders selection. While we focused on this conservative scenario, we also explored the impact of smaller deme sizes (down to 1,000 cells) (Supplementary Figs. 24 and 25). Within the model, there is no spatial partition for tumor cells within demes that proliferate via a discrete stochastic birth-and-death process (division rate p and death rate q = 1 − p; the death/birth ratio h = q/p), where the first deme is generated by the same process, beginning with a single transformed tumor cell. Simple birth–death processes on average give rise to exponential growth of each deme where the growth rate is r = ln(2p). Here we employ the following parameters: p = 0.55, q = 0.45, and thus r = ln(2 × 0.55) = 0.1 as the growth rate of deme expansion, where p and q were empirically chosen by assuming a relatively high death versus birth rate (h = q/p) = 0.82 in each cell generation, in line with previous estimates in a rapidly growing colorectal cancer metastasis (h = 0.72)50 and in early tumors (h = 0.99)51. Once the deme exceeds the maximum size, it will split into two offspring demes via sampling from a binomial distribution (Nc, p = 0.5), where Nc is the current deme size. During each cell division, the number of neutral passenger mutations that arise in the coding region of the genome follows a Poisson distribution with mean u, where an infinite sites model and constant mutation rate are assumed. Under the null model, all somatic mutations are assumed to be neutral and not confer a fitness advantage, whereas in the selection models beneficial mutations (or advantageous mutations) occur stochastically via a Poisson process with mean u during each cell division. Thus, we consider the null neutral model (s = 0) as well as varying degrees of selection: s = 0.01, 0.02, 0.03, 0.05, and 0.1, where s is the selection coefficient defined by the increase in the cell division rate when a beneficial mutation occurs in the neutral cell lineage. The cell division rate and cell death rate of a selectively beneficial clone are p_d = p × (1 + s) and q_d = 1 − p_d = 1 − p × (1 + s), respectively. The growth rate of a selective lineage within a deme is r_d = ln(2 × p_d). The parameters employed are reported in Supplementary Table 1 and include u = 1.2 within the 60 Mb of coding sequence in a diploid genome, corresponding to a mutation rate of 2 × 10^{-8} mutations per cell division per site. For the selection models, we assume u = 1 × 10^{-8} mutations per cell division for driver mutations, which is on order with the rate previously suggested by Bozic et al.53. We also investigated the impact of a lower selectively advantageous mutation rate (u_n = 1 × 10^{-8}) on the SFS, as this mimics late-arising driver mutations (Supplementary Fig. 10).

We also sought to explore how a naive model of neutral CSC-driven tumor growth would influence the resultant SFS. Here each deme comprises two subpopulations—stem cells and non-stem cells, where the stem cell fraction is p(SC). In each cell generation, stem cells divide symmetrically,
generating two stem cells with probability $\alpha$, and asymmetrically, generating one stem cell and one non-stem cell with probability $\beta$ (where $\alpha + \beta = 1$ and the probability of symmetric stem cell differentiation is thus 0). Non-stem cells can only divide with probability $\gamma$ or die with probability $\delta$ (where $\gamma + \delta = 1$). We exploit a set of parameter values, namely $\alpha = 0.15$, $\beta = 0.85$, $\gamma = 0.565$, and $\delta = 0.435$, to ensure that the maximum deme size is $\approx$ 10,000 cells and the stem cell fraction $p(\text{SC}) = 1–2\%$, consistent with estimates in solid tumors. While it is of potential interest to consider a CSC model in the context of selection, this is complicated by the need for additional parameters with little experimental support; hence, we do not investigate this here.

During virtual tumor growth, each mutation was assigned a unique index and was recorded with respect to its genealogy and host cells during the simulation, enabling analysis of its frequency in a subpopulation or the whole tumor at different stages of growth. Once the tumor reached a final size of $\approx 10^9$ cells, approximately the size when it is detectable and routinely resected, we virtually sampled one, two, four, or eight regions composed of $\approx 10^6$ cells from an individual virtual tumor (200 tumors under each of the 7 evolutionary modes, totaling 1,400 virtual tumors). The VAFs of all SSNVs in the sampled bulk subpopulation were considered the true values, whereas observed VAF values were obtained via a statistical model that mimics the random sampling of alleles during sequencing. In particular, we applied a binomial distribution ($n, f$) to generate the observed VAF of each site given its true frequency $f$ and number of covered reads $n$. The number of covered reads in each site is assumed to follow a negative binomial distribution. Here we assume depth = 80 representing 80× sequencing depth on average with a variation in parameter size of 2. A mutation is called when the number of variant reads is $\geq 2$, thereby applying the same criteria as for the actual tumors. For each virtual tumor, 100 clonal SSNVs were assigned to represent public mutations, where VAF values were simulated using the statistical model described above with mean VAF of 0.5.

Identification of subclonal SSNVs in MRS. An SSNV $m$ is defined as subclonal if all of the three following criteria are met:

1. A total probability

$$P_{m} = \prod_{i=1}^{k} p_{mi}(X_{mi} \leq S_{mi}, N_{mi} \leq f_{m}p_{mi}) < 0.05$$

where $p_{mi}$ is a binomial probability for region $i$ of observing less than or equal to $S_{mi}$ reads carrying the mutant allele out of the total reads $N_{mi}$ provided a lower bound of expected allele frequency if $m$ is public, given that the tumor purity for region $i$ equals $p_{u}$ and the total and minor copy numbers and the cellular prevalence of the SCNA where $m$ resides are equal to $n_{mi}$, $n_{bmi}$, and $p_{a_{mi}}$, respectively, within the tumor context,

$$f_{m}p_{mi} = \begin{cases} \frac{p_{u}\times n_{bmi}}{n_{mi}} & \text{if } n_{bmi} \geq 1, \text{ } n_{mi} \geq 2 \\ \left(\frac{p_{u}\times (n_{mi} - n_{bmi})}{n_{mi}}\right) & \text{otherwise} \end{cases}$$

where $n_{cma} = n_{m} \times p_{a_{mi}} \times p_{u} + 2 \times (1 - p_{a_{mi}} \times p_{u})$. For sites devoid of SCNAs, $n_{mi} = 2$, $n_{bmi} = 1$, and $p_{a_{mi}} = 0$.

2. At least one region $i$ with $\text{CCF}_{mi} \geq 95\%$ $\text{CI}_{mi} < 1$;

3. At least one region $i$ with adjusted VAF $\text{VAF}_{mi} < 0.25$. Here 0.25 was chosen because of its good performance in defining subclonality based on simulated virtual tumors (Supplementary Fig. 36).

An SSNV that does not meet one of the above criteria is considered public. SSNVs with varying patterns of LOH among regions were not included for pairwise SFS comparisons. The pooled cumulative SFS was computed when multiple samples were available. Here we employ an $f_{\text{max}}$ of 0.25 as the upper value for subclonal mutations, whereas $f_{\text{min}}$ depends on the total sequencing depth (and, hence, the number of regions sequenced) and is chosen conservatively while maximizing the inclusion of high-confidence, low-VAF SSNVs.

ITH metrics. For pairwise comparisons of regions, subclonal (private) SSNVs were defined as being either shared or region specific. Shared private SSNVs are present in both regions, whereas region-specific SSNVs are unique to one region where we reject a null model of the same VAF in the variant-missing region (given the sequencing depth) with a 5% significance level. For each pairwise SFS histogram, the bin width was optimized for visualization purposes on the basis of the number of SSNVs. Metrics capturing between-region ITH were computed for $k$ regions and

$$r = \frac{k}{2}$$

pairwise comparisons as follows:

1. \begin{align*}
\text{fHSub} &= \frac{1}{k} \sum_{i=1}^{k} \frac{\text{SM}_{i}^{\text{high}}}{\text{SM}_{i}^{\text{all}}} \\
\text{SM}_{i}^{\text{high}} &= \text{the number of high-frequency subclonal SSNVs (adjusted VAF > 0.2, hereafter referred to as VAF) and the number of all subclonal SSNVs with VAF >0.08 for region } i \\
\text{SM}_{i}^{\text{all}} &= \text{the number of all subclonal SSNVs with VAF >0.08 for region } i \\
\end{align*}

where $\text{SM}_{i}^{\text{high}}$ and $\text{SM}_{i}^{\text{all}}$ are the number of high-frequency subclonal SSNVs (adjusted VAF > 0.2, hereafter referred to as VAF) and the number of all subclonal SSNVs with VAF >0.08 for region $i$. The cutoff was set to 0.2 because above this value fHSub tends to plateau in its sensitivity to distinguishing the neutral and selection models (Supplementary Fig. 36). A lower cutoff of 0.08 was chosen empirically to satisfy the tradeoff between the number of subclonal SSNVs and variant calling errors.

2. \begin{align*}
\text{fHrs} &= \frac{1}{r} \sum_{j=1}^{r} \left( \frac{\text{RSM}_{ja}^{\text{high}}}{2 \times \text{RSM}_{ja}^{\text{all}}} + \frac{\text{RSM}_{jb}^{\text{high}}}{2 \times \text{RSM}_{jb}^{\text{all}}} \right) \\
\text{RSM}_{ja}^{\text{high}} &= \text{the number of high-frequency (VAF >0.2) region-specific SSNVs and the number of all region-specific SSNVs with VAF >0.08 for region } a, \text{ in a pairwise comparison } j \text{ between regions } a \text{ and } b \\
\text{RSM}_{ja}^{\text{all}} &= \text{the number of all region-specific SSNVs with VAF >0.08 for region } a \\
\text{RSM}_{jb}^{\text{high}} &= \text{the number of high-frequency (VAF >0.2) region-specific SSNVs and the number of all region-specific SSNVs with VAF >0.08 for region } b \\
\text{RSM}_{jb}^{\text{all}} &= \text{the number of all region-specific SSNVs with VAF >0.08 for region } b \\
\end{align*}

3. \begin{align*}
\text{FstHudson} &= \frac{1}{r} \sum_{j=1}^{r} \text{FstHudson}_{ja} \\
\text{FstHudson}_{ja} &= \frac{1}{r} \sum_{m=1}^{m} \left( f_{ja}^{m} - f_{ja}^{m} \right)^{2} - \frac{f_{ja}^{m} \times \left(1 - f_{ja}^{m}\right)}{d_{m}^{2} - 1} \\
&= \frac{1}{r} \sum_{m=1}^{m} \left( f_{ja}^{m} - f_{ja}^{m} \right)^{2} - \frac{f_{ja}^{m} \times \left(1 - f_{ja}^{m}\right)}{d_{m}^{2} - 1} \\
\text{where } f_{ja}^{m} &= \text{the VAF for SSNV } m \text{ and } d_{m}^{2} &= \text{the sequencing depth for SSNV } m \text{ in region } a. \text{ The genetic variance components (numerator and denominator) are averaged separately to obtain a ratio combining the Hudson FST estimates across all } m \text{ SSNVs.} \\
\end{align*}

4. \begin{align*}
\text{Ksd} &= \frac{1}{r} \sum_{j=1}^{r} \text{Ksd}_{j} \\
\text{Ksd}_{j} &= \text{max}\left| F_{a} - F_{b} \right| \\
\text{where } F_{a} &= \text{the cumulative SFS of region } a, \text{ in a pairwise comparison } j \text{ between regions } a \text{ and } b \\
\end{align*}

5. \begin{align*}
\text{rAuc} &= \frac{\text{Auc}_{\text{merged}}}{\text{Auc}_{\text{theoretical}}} \\
\text{where } \text{Auc}_{\text{merged}} &= \text{the AUC for merged reads} \\
\text{Auc}_{\text{theoretical}} &= \text{the AUC for theoretical reads} \\
\end{align*}
corresponding to the ratio of the area under the pooled cumulative SFS to the area under a theoretical cumulative SFS assuming neutral exponential growth of a well-mixed population\(^{21,25}\). For MRS, the pooled VAF is the total number of alternative alleles divided by the total read depth. As this represents the alternative allele frequency pooled across tumor regions, it should capture overall tumor dynamics but not between-region diversity and it complements other ITH metrics.

To evaluate the power (at a significance level of 0.10) or sensitivity of ITH metrics to distinguish a specific alternate model from the neutral model in the simulated data given varying numbers of samples \((n = 1, 2, 4, \text{ and } 8)\) or variable sequencing depths for a single sample (80–640×), we employed \(r\)AUC, as it is applicable to single-sample data, as well as \(f\)Hub, one of the MRS-specific statistics. The power was computed empirically as the percentage of virtual tumors under an alternative model for which the statistic \((r\text{AUC} \text{ or } f\text{Hub})\) was greater than 95% or less than 5% of the corresponding statistic in the neutral model (taking the larger percentage).

**Evolutionary mode classifier.** A radial basis function (RBF) kernel SVM was built on the basis of 1,400 simulated tumors derived from seven growth models (200 for each of neutral, neutral CSC, and \(s = 0.01, 0.02, 0.03, 0.05, \text{ and } 0.1\)). We grouped virtual tumors simulated under the neutral, neutral CSC, and \(s = 0.01\) models as ‘effectively neutral’ and those simulated under higher selection coefficients (\(\text{CSC, and } s > 0.01\)) as ‘selection models’ and evaluated the different combinations of these models using a second SVM. To evaluate the relative importance of different combinations of the five ITH metrics for classification, SVMs were run 20 times for each of 26 possible combinations of five statistics with the same seed used for random splitting, where four of the five virtual tumors were used for training and one was used for testing, and the resulting ROC AUCs were compared (Supplementary Fig. 27). An SVM was also built using the two major ICs obtained from ICA of the five ITH metrics where the decision boundaries are shown on the ICA scatterplots. ICA was performed on features derived from the virtual tumors and patient tumors for \(n = 2\) (Supplementary Fig. 28), \(n = 4\) (Fig. 6b), and \(n = 8\) (Supplementary Fig. 29) virtual tumor regions. The performance of the SVM in distinguishing each alternative model from the neutral model was evaluated by comparing 100 virtual tumors for training and 100 virtual tumors for testing (Supplementary Fig. 26).

**Functionality assessment of private and public SSNVs.** The ratio of private SSNVs at more functional (MF) relative to less functional (LF) sites was determined as previously described\(^{46}\) to evaluate the correlation between dMF/dLF and various ITH metrics derived from the SFS. SSNVs were considered more functional if classified by PolyPhen-2 as ‘damaging’ or ‘possibly damaging’ and less functional if classified as ‘benign’. The dMF/dLF ratio was calculated by normalizing the MF/LF ratio for private SSNVs in each tumor to a background MF/LF ratio based on random substitutions in the mutated genes. We also determined the fold enrichment for driver genes (defined on the basis of IntOGen v.2016.5) among non-silent public SSNVs and the correlation with various ITH metrics.

**Code availability.** Code for the simulation studies and the variant assurance pipeline (VAP) is available at [https://github.com/cancersysbio/VirtualTumorEvolution](https://github.com/cancersysbio/VirtualTumorEvolution) and [https://github.com/cancersysbio/VAP](https://github.com/cancersysbio/VAP).

**Data availability.** The single-gland WES data and xenograft WES data are available at EMBL-EBI ArrayExpress under accession E-MTAB-5547. Data from previously published studies are available at the European Genome-phenome Archive (EGA) under accessions EGAD00001000900, EGAD00001000984, and EGAD00001000113.

---

47. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
48. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 24, 1881–1893 (2014).
49. Waclaw, B. et al. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature 525, 261–264 (2015).
50. Diaz, L.A. Jr. et al. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature 486, 537–540 (2012).
51. Bozic, I. et al. Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. USA 107, 18545–18550 (2010).
52. Visvader, J.E. & Lindeman, G.J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat. Rev. Cancer 8, 755–768 (2008).
53. Wand, M.P. Data-based choice of histogram bin width. Am. Stat. 51, 59 (1997).
54. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting \(F_{ST}\): the impact of rare variants. Genome Res. 23, 1514–1521 (2013).
55. Kuhn, M. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26 (2008).