nature genetics

# A Big Bang model of human colorectal tumor growth

Andrea Sottoriva[1,6], Haeyoun Kang[2,3], Zhicheng Ma[1,6], Trevor A Graham[4,5], Matthew P Salomon[1], Junsong Zhao[1], Paul Marjoram[1], Kimberly Siegmund[1], Michael F Press[2], Darryl Shibata[2] & Christina Curtis[1,6]

**What happens in early, still undetectable human malignancies is unknown because direct observations are impractical. Here we present and validate a 'Big Bang' model, whereby tumors grow predominantly as a single expansion producing numerous intermixed subclones that are not subject to stringent selection and where both public (clonal) and most detectable private (subclonal) alterations arise early during growth. Genomic profiling of 349 individual glands from 15 colorectal tumors showed an absence of selective sweeps, uniformly high intratumoral heterogeneity (ITH) and subclone mixing in distant regions, as postulated by our model. We also verified the prediction that most detectable ITH originates from early private alterations and not from later clonal expansions, thus exposing the profile of the primordial tumor. Moreover, some tumors appear 'born to be bad', with subclone mixing indicative of early malignant potential. This new model provides a quantitative framework to interpret tumor growth dynamics and the origins of ITH, with important clinical implications.**





The growth of human malignancies cannot be directly observed. In particular, the earliest events in the growth of a large tumor are unknown. What happens during these first cell divisions may provide clues as to how to better prevent, detect and treat cancers. Because tumor growth is an evolutionary process and the ancestral history is recorded within tumor cell genomes[1–3], detailed information on the early growth phase may be encoded in the patterns of genomic ITH present in the final neoplasm. Specifically, in the absence of selective sweeps, it is feasible to recover the genomic profile of the primordial tumor. This task is possible because private (subclonal) alterations, including copy number aberrations (CNAs) and point mutations, that occur early during growth should be 'pervasive' in the final neoplasm, where pervasive refers to private alterations that are found throughout the tumor but are not dominant. Experimentally, pervasive alterations can be detected through systematic sampling and genomic profiling of numerous regions of the same neoplasm. The initial events in neoplastic transformation are thought to occur through the stepwise accumulation of driver alterations[4], whereas the growth dynamics of established neoplasms remain poorly characterized. In particular, extensive ITH and branching phylogenies identified by cancer genomic studies[5–9] suggest that the same linear paradigm does not apply to the subsequent growth of established tumors, such as colorectal carcinomas and advanced adenomas. However, the origins of ITH are unknown, and a quantitative framework to describe the dynamics of tumor growth is needed.

Here we propose a Big Bang model where, after the initial transformation, colorectal tumors grow predominantly as a single expansion populated by numerous intermixed subclones (**Fig. 1a**). As expected, public alterations in the initiating cell will be present in all tumor cells (clonal). In contrast, although new private alterations will continuously be generated as a result of replication errors, only the earliest will be pervasive, whereas later alterations will be localized in progressively smaller tumor subpopulations. Although private alterations acquired during growth may confer survival advantages, selective sweeps that substantially alter the clonal composition of the final tumor are predicted to be extremely rare owing to the rapidly expanding population and spatial constraints[10–12]. Hence, the timing of an alteration rather than clonal selection for that alteration is the primary determinant of its pervasiveness. Notably, most observable private alterations that give rise to ITH are generated early after the transition to an advanced tumor, well before the neoplasm becomes clinically detectable. Given the absence of sequential selective sweeps, our model anticipates uniformly high levels of ITH throughout the neoplasm. Moreover, in some tumors, early subclone mixing followed by scattering to different distant tumor regions might occur (for example, **Fig. 1a**, red subclone). This phenomenon results in variegated tumor cell populations, where the spatial relationship between cells does not necessarily recapitulate their clonal relationship.

An example of the variegation predicted by the Big Bang model is shown in **Figure 1b**. Progeny of the first initiating tumor cell propagate public alterations but also acquire new private alterations (colored areas), resulting in ITH within the newly formed small, primordial tumor, which can subsequently scatter to distant regions during growth. For instance, the earliest alterations (shown in red) can be scattered to opposite sides of the neoplasm during tumor expansion, despite remaining private and non-dominant. This mechanism

[1]Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA. [2]Department of Pathology, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA. [3]Department of Pathology, CHA University, Seongnam-si, South Korea. [4]Center for Evolution and Cancer, University of California, San Francisco, San Francisco, California, USA. [5]Centre for Tumor Biology, Barts Cancer Institute, Queen Mary University of London, London, UK. [6]Present addresses: Division of Molecular Pathology, The Institute of Cancer Research, London, UK (A.S.), Department of Medicine, Stanford University, Stanford, California, USA (Z.M. and C.C.) and Department of Genetics, Stanford University, Stanford, California, USA (Z.M. and C.C.). Correspondence should be addressed to D.S. (dshibata@usc.edu) or C.C. (cncurtis@stanford.edu).

Received 12 October 2014; accepted 12 January 2015; published online 9 February 2015; doi:10.1038/ng.3214

generates patterns of genetic variegation in the tumor. Therefore, clones harboring early private alterations (red or yellow) will be more pervasive in the final tumor, whereas late-arising clones will not have had time to expand to a detectable size, regardless of their relative fitness advantage (pink, black, green and blue). This simple model predicts that early private alterations underlie the extensive ITH commonly detected in human neoplasms. Hence, public as well as the majority of detectable private alterations occur early during tumor growth.

Here we experimentally evaluate the predictions of the Big Bang tumor model by profiling 349 individual tumor glands sampled from opposite sides (arbitrarily defined as 'right' and 'left') of 15 colorectal carcinomas and large adenomas (**Supplementary Table 1**) using orthogonal genomic techniques—namely, whole-genome array-based profiling of CNAs, whole-exome sequencing, targeted deep sequencing, FISH and neutral methylation tag sequencing. By analyzing single tumor glands composed of <10,000 cells, this approach enables the detection of alterations that occur in a fraction of tumor cells with remarkable sensitivity. At this level of resolution, we find unexpected spatial structure, indicative of order amid the apparent chaos of genomic ITH. By integrating these data in a robust statistical inference framework based on a spatial computational model of tumor growth, we also verified that most ITH detectable with current technologies arises early during tumor growth and that the genomic profile of the primordial tumor can be recovered from the present-day neoplasm.

## RESULTS

### Sampling individual tumor glands

Colorectal cancer (CRC) represents an optimal system in which to study the dynamics of tumor growth as both the normal and neoplastic colon are organized into glandular epithelial structures, where neighboring cells within a gland share a recent common ancestry[13] and microenvironment, with gland fission being the primary mode of growth[14,15]. Here we systematically sampled an average of 23 individual tumor glands and 2 'bulk' fragments from the right and left sides (**Fig. 1c**) of 4 large, mitotically advanced adenomas and 11 carcinomas (**Supplementary Table 1**), totaling 349 tumor glands and 22 bulk samples. This approach enables the highly sensitive detection of subclonal alterations (<10,000 cells per gland out of 100 billion cells in a tumor; 0.00001%).

### Single-gland copy number profiles show variegation

Copy number profiles can be used to reconstruct tumor phylogenies[6,8,16], and, by profiling single glands, it is possible to do so with unprecedented accuracy. We exploited whole-genome SNP array–based copy number data derived from individual glands (7–10 per tumor; $n$ = 127 total), left and right bulk tumor fragments (>3 cm apart) and corresponding matched normal tissues to systematically evaluate the spatial distribution of CNAs throughout each tumor. These data showed striking spatial patterns, which were classified as follows: (i) public (found in all glands of the tumor); (ii) private, side specific (found in all glands from one tumor side only); (iii) private, side variegated (found in all glands from one tumor side and in some glands from the opposite side); (iv) private, variegated (found in a subset of glands from both sides); (v) private, regional (found in more than one but not all glands from one tumor side only); and (vi) private, unique (found in a single gland).

Consistent with their likely monoclonal origin from a single aberrant colon crypt[17], most tumors exhibited public alterations acquired before initiation that were present in all glands (**Fig. 2a**, **Supplementary Fig. 1a** and **Supplementary Table 2**). Adenomas were more chromosomally stable and less genomically complex than carcinomas, despite their comparably large size (**Supplementary Table 1**). Adenomas were characterized by side-specific and unique CNAs that clearly segregated between tumor sides. In contrast, the majority of carcinomas (M, N, O, U, CA, CO and R) exhibited the same private CNA in individual glands from opposite sides of the tumor (variegated and/or side variegated), as reflected in the underlying phylogenetic trees (**Fig. 2b** and **Supplementary Fig. 2**). This corresponds to the patterns of variegation presented in **Figure 1b** where an early private alteration originating in the primordial tumor is scattered to distant tumor sites and appears pervasive in the neoplasm, despite remaining subclonal.

Such genetic variegation has been noted in leukemia[18] and solid tumors[19,20] but is often obscured by the prevailing approach of analyzing bulk tissue rather than individual glands or cells. To verify that the individual glands analyzed were representative of the larger tumor mass, we profiled the right and left bulk tumor fragments (**Fig. 2a** and **Supplementary Fig. 1a**, bulk tracks: LB, left; RB, right). We found that 99% of the non-unique CNAs present in the glands were also present in the bulk tumor fragments and that the majority of the private CNAs identified in glands were present as a mixture in the

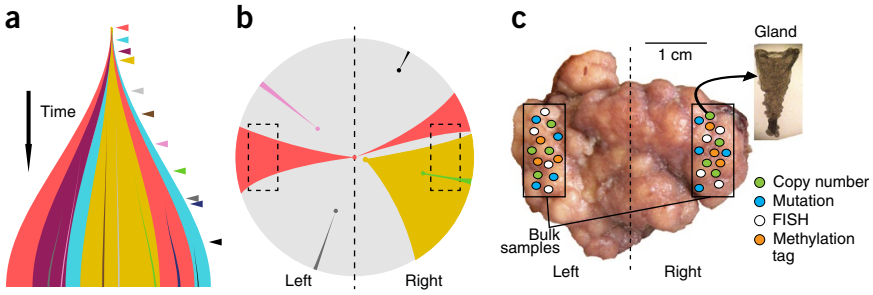


**Figure 1** The Big Bang model of tumor growth. (**a**) After initiation, a tumor grows predominantly as a single expansion populated by numerous heterogeneous subclones. ITH results from private alterations (colored arrowheads) that continuously accumulate owing to replication errors. In addition to public alterations present in the first transformed cell, private alterations acquired early persist and become pervasive in the final tumor although remaining non-dominant (colored segments). Late-arising alterations are only present in small regions of the tumor. (**b**) In the Big Bang model, the pervasiveness of private alterations depends on when the alteration occurs during growth, rather than on selection for that alteration. The schematic illustrates how early private alterations, despite remaining non-dominant, are pervasive within the tumor (for example, red and yellow) and can be found in distant regions, thus appearing variegated (for example, red). This is owing to aberrant subclone mixing in the primordial tumor, followed by scattering during expansion. Late alterations are restricted to small regions (for example, black, pink, gray) and are essentially undetectable by conventional bulk genomic profiling. Distance from the dashed vertical axis corresponds to increasingly late onset for alterations. Dashed boxes represent sampled regions. (**c**) We sampled an average of 23 individual tumor glands (<10,000 cells) from distant regions (~0.5 $cm^3$ in size) and bulk (left and right) samples from the remaining tissue. Samples were profiled using several genomic techniques, including copy number analysis, whole-exome and targeted sequencing, neutral methylation tag sequencing and FISH, providing a panoramic view of genomic alterations throughout the tumor on multiple spatial scales.

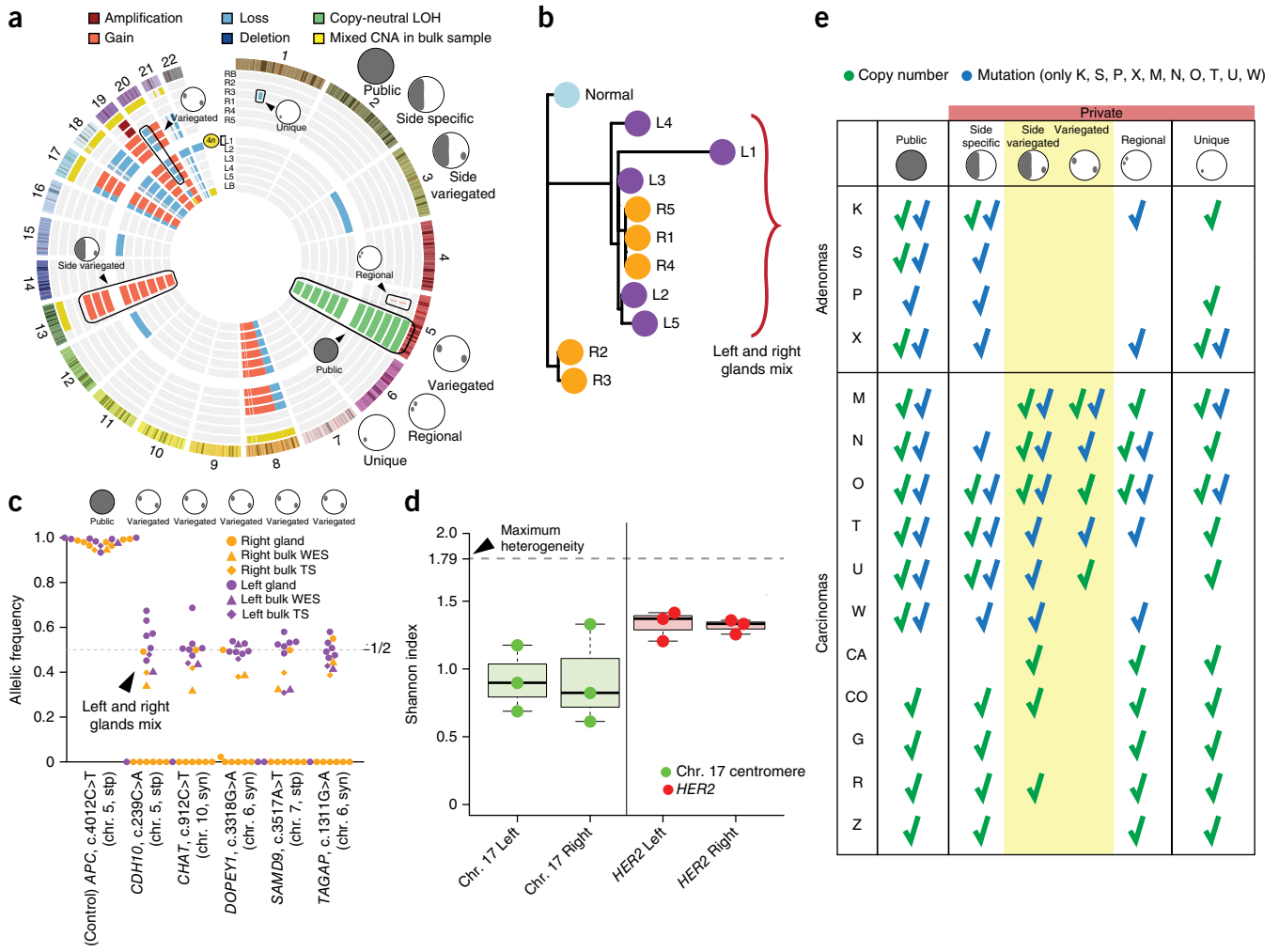


**Figure 2** The spatial distribution of ITH shows subclone mixing and the absence of clonal expansions. (**a**) Circos plot representation of CNAs in individual glands and bulk samples for carcinoma M (shown throughout this figure). LOH, loss of heterozygosity; L, left; R, right. (**b**) Gland-level copy number analysis was employed to reconstruct the tumor phylogeny. Mixing of glands from opposite regions is apparent, where glands from left and right tumor regions are colored purple and orange, respectively. (**c**) Targeted sequencing of patient-specific mutations in individual glands showed variegation in subsets of glands from opposite sides, thus confirming subclone mixing at the mutational level. A public *APC* mutation is shown as a clonal control (with LOH noted on chromosome 5). WES, whole-exome sequencing; TS, targeted sequencing; syn, synonomous; stp, stop gain. (**d**) FISH performed using *HER2* probes (red) and corresponding chromosome 17 centromeric probes (green) showed high variability in copy number states between cells within a gland, as summarized by the Shannon index. For each group, box plots show the median, limited by the 25th (Q1) and 75th (Q3) percentiles, where whiskers represent the extremes of the maximum or Q3 + 1.5(Q3 – Q1) and the minimum or Q1 – 1.5(Q3 – Q1). The maximum possible ITH value ("maximum heterogeneity") corresponds to an index of 1.79 (99% of the FISH counts; range of 0–5). (**e**) Summary of the characteristic spatial patterns and types of alterations in each tumor. Whereas adenomas were characterized by low chromosomal instability and the segregation of alterations, carcinomas harbored side-variegated and variegated alterations (7/11 at the copy number level and 6/6 at the mutational level). Yellow shading highlights variegated alterations.

respective bulk fragment (**Supplementary Fig. 3**). All CNAs evident in a bulk sample were also detected in one or more corresponding tumor glands. Moreover, we emphasize that, if we had sampled only a portion of the tumor (for example, only the right side or only the left side), we would have reconstructed erroneous phylogenies, as demonstrated in **Supplementary Figure 4** and as noted by others[21]. Although unlikely, we cannot exclude the possibility that the same CNA could arise independently in different glands. Hence, we also evaluated variegation at the mutational level.

## Single-gland sequencing confirms variegation

To examine mutational heterogeneity, we performed whole-exome sequencing of the bulk tumor samples (left and right sides) and adjacent normal tissue from each of the adenomas and for carcinomas M, N, O, T, U and W. On the basis of the spectrum of somatic mutations present in each bulk tumor sample, we selected a panel of patient-specific private mutations and known drivers of CRC for deep targeted sequencing (>600× mean target coverage) in individual glands ($n$ = 102) and the respective bulk tumor fragments ($n$ = 20).

All sequenced tumors, except for adenoma S and carcinomas O and W (with microsatellite instability (MSI)), harbored public nonsense mutations in *APC*. Public missense mutations in *KRAS* were found in samples N, P, S, W and X, whereas public missense *TP53* mutations were only found in carcinomas (M, N, and T), as previously reported[4]. Notably, the mutational data corroborated the findings at the CNA level, providing further evidence for the striking segregation

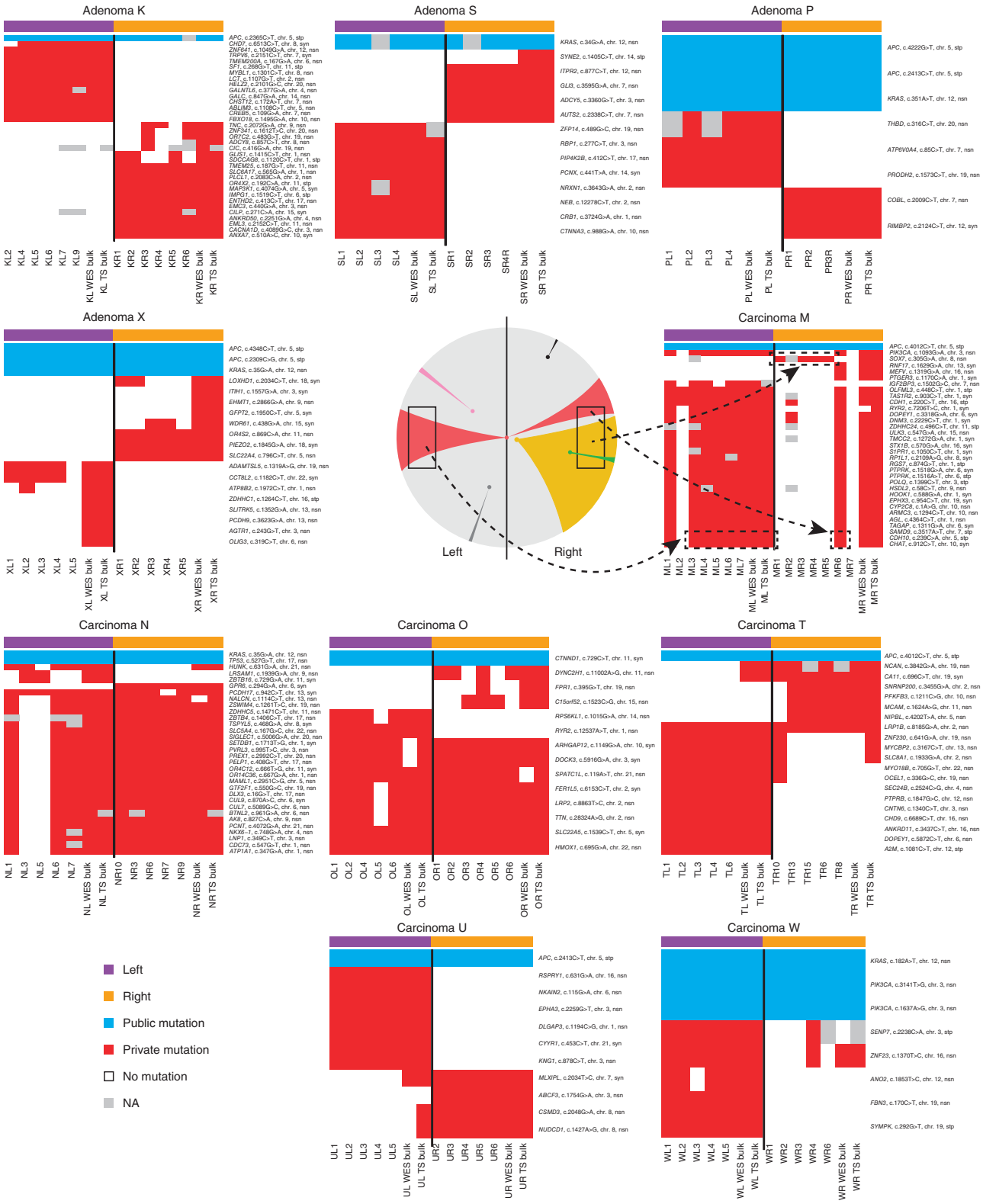


**Figure 3** Single-gland targeted sequencing confirms the predictions of the Big Bang model and exposes variegation in carcinomas but not adenomas. Heat maps indicate the presence of representative public and private mutations across multiple individual glands per tumor, where targeted sequencing (TS) and whole-exome sequencing (WES) of the bulk tumor is included for comparison. In all the adenomas, private mutations are confined to a single tumor side (regional and side-specific events), whereas, in invasive carcinomas, the same private mutation is found in distant regions of the neoplasm, despite remaining non-dominant. These patterns of genetic variegation are indicative of subclone mixing in the early neoplasm followed by scattering. For representative carcinoma M, the mutational data are summarized according to the schematic in **Figure 1b**, where variegated mutations (red) occurred early and scattered to distant tumor regions. Regional mutations (yellow) occurred later and were confined to smaller regions of the neoplasm. Heat maps indicate the presence of representative public and private mutations (nsn, nonsynonymous; syn, synonymous; stp, stop gain; NA, not available) across multiple individual glands per tumor.

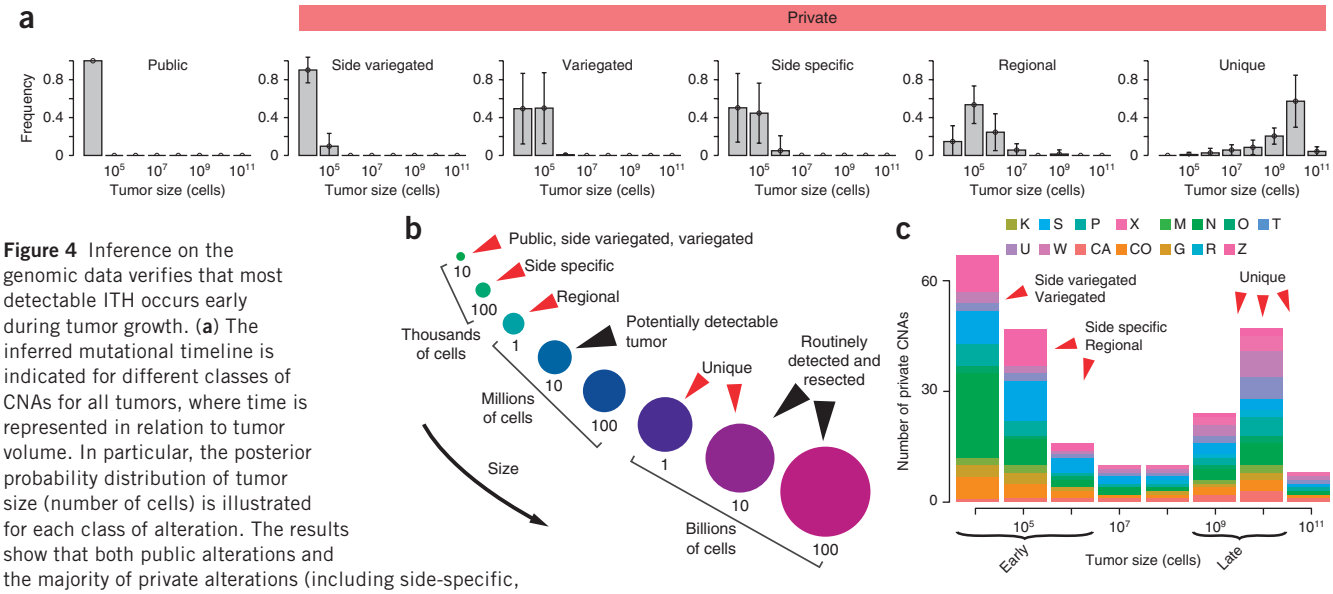


**Figure 4** Inference on the genomic data verifies that most detectable ITH occurs early during tumor growth. (**a**) The inferred mutational timeline is indicated for different classes of CNAs for all tumors, where time is represented in relation to tumor volume. In particular, the posterior probability distribution of tumor size (number of cells) is illustrated for each class of alteration. The results show that both public alterations and the majority of private alterations (including side-specific, side-variegated and variegated events) occur very early after the transition to an advanced neoplasm, when the tumor is composed of fewer than 1 million cells, whereas unique mutations occur late. Error bars, s.d. (**b**) A schematic of the mutational timeline (from **a**) illustrates that the majority of detectable non-unique alterations occur early, when the tumor is orders of magnitude smaller than can be clinically detected. As reliable estimates of cell cycle duration are not available and somatic alterations depend on cell division rates rather than time, tumor size is used as a surrogate for time. (**c**) By applying the inferred mutational timeline to the whole-genome CNA profiles for each patient, it is apparent that early CNAs dominate the genomic landscape. Here early events correspond to alterations that took place when the tumor had $<1 \times 10^6$ cells and late alterations correspond to those that occurred after the tumor reached $1 \times 10^9$ cells. For simplicity, only private alterations are represented, as all public alterations occur early.

of subclones in all adenomas, whereas variegation, indicative of early subclone mixing, was observed in the carcinomas. A summary of the characteristic spatial patterns in each tumor is reported in **Figure 2e**. Here variegation, determined on the basis of the presence of the same single-nucleotide variant (SNV) in glands from distant tumor sides, was found in all carcinomas (**Fig. 2c** and **Supplementary Figs. 1b** and **5**), despite the bias against detecting this phenomenon due to only 7–10 glands being profiled per tumor.

The targeted sequencing results for private mutations (red) and representative public mutations (blue) are presented in **Figure 3**. As shown for carcinoma M (**Figs. 2** and **3**), mutations in *SAMD9*, *CDH10* and *CHAT* were variegated and recapitulate the predictions of the Big Bang model (**Fig. 1b**), where a private mutation originates in the primordial tumor and subsequently scatters as a result of expansion. In contrast, early public mutations in *APC* were found in all cells in the neoplasm and represent a clonal control. Private mutations detectable in the bulk specimens were always present in at least one of the sampled glands, consistent with the pervasive nature of ITH. In addition, within small gland populations, private mutations will eventually be lost or fixed. Private mutations were clonal within the gland, reflecting their early acquisition and sufficient time for loss or fixation via cell turnover or neutral drift[22].

Hypothetically, glands harboring the same private mutations found on opposite tumor sides (several centimeters apart) could result from alternative mechanisms such as late-arising mutations and subsequent migration or tumor cell reseeding[23]. However, such migration is unlikely because the private mutations were clonal within individual glands, and the migration of whole glands is improbable. Instead, subclone mixing is efficient in an early, small malignancy characterized by loss of normal cell adhesion and disorganized growth. The ensuing expansion allows early private mutations to become fixed within glands, pervasive in the tumor and scattered to opposite tumor sides, thus generating patterns of variegation. Indeed, variegation was restricted to carcinomas (**Figs. 2e** and **3**). This observation suggests that certain malignant features, such as abnormal mobility, might be expressed very early, even before visible invasion and/or metastasis occurs, implying that some tumors are 'born to be bad'. An illustrative simulation demonstrates that subclone mixing in an early tumor followed by expansion can create complex patterns of variegation (**Supplementary Fig. 6**). In contrast, when the same mutation arises later, subclones appear segregated, irrespective of their relative fitness advantage.

## Single-cell profiling shows uniformly high ITH

The fixation of private alterations within a gland could occur through stepwise selection, where cells with even a slight selective advantage will sweep through the gland. In this scenario, there should be very little within-gland heterogeneity. By contrast, a single Big Bang expansion implies that individual glands in the final tumor are relatively old populations that should exhibit similar within-gland diversity. We evaluated copy number heterogeneity between physically adjacent single cells by FISH in a subset of tumor glands ($n = 65$) and adjacent normal glands ($n = 22$). In particular, we assayed for *HER2* (*ERBB2*) gene amplification, a driver event in breast and gastric cancers, which has been implicated in CRC[24]. These data showed a high degree of variability in copy number between physically adjacent cells within the same gland as quantified by the Shannon index[19]. Notably, this diversity was uniformly high throughout the tumor (**Fig. 2d**, **Supplementary Fig. 1c** and **Supplementary Table 3**). Because alterations should fix quickly within small populations[25], this finding suggests the absence of recent clonal expansions within glands. Variation in copy number between nearby cells is reportedly common in CRC owing to chromosomal instability (CIN)[26] and may be important for tumor initiation[27] and progression[28]. Moreover, it can be used to assess genetic and phenotypic diversity in response to chemotherapy[29].

We also evaluated epigenetic passenger mutations through ultra-deep single-molecule methylation tag sequencing of individual glands

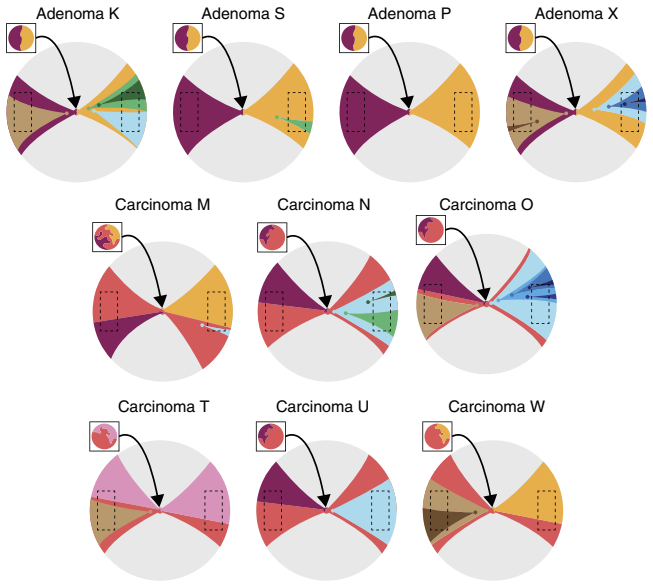


**Figure 5** Schematic of spatiotemporal Big Bang growth dynamics. For each tumor profiled at the mutational level, the phylogeny was reconstructed from the single-gland and bulk tumor data (Online Methods) to define subclones. The relative timing during which each subclone arose was specified on the basis of the inferred mutational timeline (**Fig. 4a** and **Supplementary Fig. 9c**) for the different classes of private alteration (variegated, side specific, regional and unique). By combining information on the mutational timeline and tumor subclonal architecture, we can approximately reconstruct patient-specific spatiotemporal evolutionary dynamics, as shown in this schematic. The topographical distribution of different subclones is illustrated by distinct colors, and distance from the tumor origin (arrowhead) corresponds to the increasingly late onset of alterations. Variegated and side-variegated subclones occurred very early within the primordial tumor (<1 million cells) and are shown within the inset square representing a magnified view of the primordial neoplasm. Regional and unique subclones arose later and are represented outside the inset square. Dashed boxes represent the regions of the tumor that were experimentally sampled. This schematic shows how, in the Big Bang tumor model, the prevalence of a private mutation depends on when it arose during tumor expansion, rather than on selection for that mutation. The schematic also illustrates that, although all tumors exhibit Big Bang dynamics, subclone mixing is restricted to carcinomas, whereas adenomas are characterized by subclone segregation.



($n = 55$), which provides an efficient means to infer cell ancestries in normal[30,31] and cancerous[5,9] tissues. These data showed uniformly high ITH (**Supplementary Fig. 1d**), reflecting similar tumor age in different glands and opposite sides of the neoplasm, in agreement with the FISH analysis. In particular, numerous mitotic subclones within the same gland were found in the majority (49/55) of samples (**Supplementary Fig. 1e**), supporting the absence of recent selective sweeps, as predicted by the Big Bang model.

### Statistical inference verifies the Big Bang model predictions

The most striking prediction of the Big Bang tumor model is that, even though new alterations occur continuously throughout tumor growth, the majority of private alterations that can be detected occur early after the transition to an advanced tumor, rather than as a result of the subsequent selection of *de novo* clones. To quantitatively test this prediction, we extended our previously described statistical inference framework approach[9] to take as input copy number and mutational data from multiple tumor glands and to account for differences in subclone fitness and contributions from the local microenvironment. The framework uses Approximate Bayesian Computation (ABC)[32] and three-dimensional mathematical modeling to infer patient-specific tumor characteristics, including the mutation rate, subclone fitness changes and the mutational timeline, given the observed multiple-sampling genomic data (**Supplementary Fig. 7**). The model simulates the expansion of a tumor containing ~8 million glands, corresponding to a realistically sized neoplasm composed of ~80 billion cells with a diameter of ~5.3 cm, and accounts for gland proliferation in three-dimensional space, somatic alterations (CNAs and point mutations) and changes in subclone fitness (see the Online Methods for details).

The inference results indicate that, although changes in subclone fitness can be detected (**Supplementary Fig. 8a**), their effects on the clonal composition of the tumor are limited, as corroborated by the presence of adjacent glands with different fitness (**Supplementary Fig. 8b**). The magnitude of fitness changes was variable in carcinomas, whereas adenomas exhibited limited or no differences in fitness between subclones. Mutation rates were also elevated in carcinomas ($1 \times 10^{-6}$ to $1 \times 10^{-5}$) as compared to adenomas ($1 \times 10^{-6}$ alterations per division) (**Supplementary Fig. 8a**), similarly highlighting within-tumor variability in clonal dynamics and key phenotypic differences between adenomas and carcinomas. We also employed this framework to infer the timeline during which different classes of alteration occur and quantitatively show that, for each of the tumors assayed, both public and most private alterations (side specific, side variegated and variegated) occurred early (**Fig. 4a**), when the malignancy had fewer than 10,000–100,000 cells (**Fig. 4b**), where size is used as a surrogate for tumor age. This is approximately 100–1,000 times smaller than the size at which colorectal tumors are potentially detectable (~1 $mm^3$ or $1 \times 10^6$ cells) and 1 million times smaller than is typical at the time of surgical resection (the source of sampled tissue). Even regional alterations tended to occur before the tumor would be clinically detectable, whereas unique alterations arose later, as expected. These findings hold irrespective of tumor-specific characteristics. The same conclusions were obtained using mutational data as input to the framework (**Supplementary Fig. 9**). By organizing the observed patient-level genomic profiles according to the inferred mutational timeline, it is evident that early subclonal alterations dominate the genomic landscape (**Fig. 4c**).

Using single-gland and bulk tumor mutational profiles (**Fig. 3**), we reconstructed tumor phylogenies (Online Methods) to define subclones, or groups of glands harboring the same private mutations. By superimposing the inferred mutational timelines for different classes of alterations (**Fig. 4a** and **Supplementary Fig. 9c**), we then determined the relative timing at which each subclone arose. This allows for the approximate reconstruction of patient-specific spatiotemporal evolutionary dynamics, as depicted schematically in **Figure 5**, and shows that the pervasiveness of a private mutation depends on when it arose during expansion, rather than as a result of selection for that mutation. This schematic also illustrates that, whereas all tumors exhibit Big Bang dynamics, early subclone mixing in the primordial tumor is restricted to carcinomas.

Clonal heterogeneity could alternatively be due to distinct local microenvironmental niches within the neoplasm that select for clones with different genomic profiles[1]. To investigate this scenario, we introduced microenvironmental niches in our model (Online Methods and **Supplementary Fig. 10**). The inferred parameters were in agreement with the results from the microenvironment-free model for both CNAs and mutations (**Supplementary Fig. 11**), further supporting our conclusions. This follows from the fact that microenvironmental selection acts passively on existing variation. Of note,

we have modeled the microenvironment as a static entity and do not account for the possibility that tumor cells might dynamically alter their environment, although such mechanisms may have a role in later growth[3]. In the future, it will be of interest to examine more complex interactions between cells and their microenvironment, as well as to measure interclonal interactions, which have recently been described in breast cancer[33,34].

## DISCUSSION

Tumor initiation is characterized by the sequential stepwise accumulation of alterations, leading to the expansion of clones with selective growth advantages, such that the fittest clone eventually dominates[4]. The sequential model of colorectal tumorigenesis is corroborated by epidemiological data on CRC incidence[35]. This model has often been postulated to describe the subsequent growth of an established tumor. In this scenario, further growth within an advanced tumor results from the acquisition of new driver mutations followed by selective sweeps and large clonal expansions. Within this model, ITH represents a transitory state between selective sweeps. As this model implies the occurrence of multiple sweeps, numerous drivers of tumor growth are anticipated. However, relatively few putative driver mutations have been identified in individual tumors[36].

Recent studies in primary CRCs indicate that selective sweeps and large clonal expansions are infrequent after transformation[13,37,38] and predict star-shaped phylogenies[13,37]. Studies in other cancers similarly highlight such branched phylogenies[39] and punctuated clonal evolution[6,40]. Moreover, karyotypic chaos[41], stress-induced mutational bursts[42] and chromothrypsis[43], a cataclysmic event involving surges of chromosomal rearrangements, have been reported. Evidently, sequential clonal evolution does not accurately describe the patterns of ITH found in human cancers.

Here we propose and test the predictions of a Big Bang model, whereby, as a result of a single clonal expansion, most detectable ITH occurs early after the transition to an advanced tumor. In this model, owing to constraints on clonal selection, early private alterations are pervasive in the final neoplasm, despite remaining non-dominant. Indeed, only very strongly advantageous mutations are likely to be fixed in realistic time scales[12] within rapidly expanding populations, where spatial structure delays the expansion of an advantageous mutation[10–12]. Such spatial constraints in solid tumors[1,13,44] underline the limits with which selective forces drive tumor expansion. Hence, both public and the majority of detectable private alterations occur early during tumor growth. Although private alterations continuously occur, only those that occur early have time for the corresponding clone to expand to a detectable size. The Big Bang model explains why ITH is pervasive in human tumors and provides a theoretical framework to describe the underlying clonal dynamics. The star-shaped phylogenies predicted by the Big Bang model are also compatible with the long-lived lineages of the cancer stem cell model[45], wherein a malignancy is driven by a small number of self-renewing cells. We demonstrate that Big Bang dynamics are robust to changes in subclone fitness and local microenvironment, which might explain why they are observed in many tumors.

The Big Bang model explains many poorly understood features of cancer genomic data, with the following implications: (i) ITH is an inherent characteristic of colorectal tumors that arises early and continuously increases during growth, and it is not significantly constrained by clonal selection; branched phylogenies naturally follow from the Big Bang model; (ii) substantial clonal expansions or selective sweeps are extremely rare after the transition to an advanced tumor owing to the dynamics and spatial constraints of the rapidly growing population and the formation of microenvironmental niches; (iii) both public and the majority of detectable private alterations arise early and become pervasive during tumor growth, thereby dominating the genomic structure of the neoplasm; and (iv) potentially aggressive subclones may remain rare or even undetectable in the primary tumor, despite their relative fitness advantage, providing a heterogeneous substrate to fuel resistance in response to selective pressures from treatment.

A number of clinical implications also follow from the Big Bang model. For example, it is uncertain why certain large tumors remain localized, whereas others eventually invade and metastasize. Variegated alterations were found in the majority of invasive carcinomas but in none of the adenomas. Hence, variegation might reflect the early expression of an invasive phenotype (abnormal cell intermixing), such that some tumors are 'born to be bad'. In other words, malignant potential is determined early, as previously proposed[46,47]. Moreover, the degree of subclone mixing might be a readout of subsequent invasiveness and could represent a new biomarker for predicting which adenomas will become invasive versus remain indolent. Another clinical implication that follows from the timing of mutation being the primary determinant of whether a subclone is pervasive in a tumor is that 'dangerous' treatment-resistant clones that occur late will be undetectable, presenting obvious challenges for personalized medicine. This is in line with recent reports that minor cell subpopulations can drive tumor growth[34] and with the presence of preexisting, intrinsically resistant subclones that contribute to poor treatment response[48].

Not every tumor may exhibit Big Bang dynamics, and 'selective bottlenecks' may be common for markedly different environments, such as in the context of metastatic seeding to foreign sites or during treatment. However, for primary tumors that arise predominantly as single clonal expansions, this new model represents a theoretical framework in which to interpret cancer genomic data and predicts that the earliest events should be pervasive in the final neoplasm. This concept shares an interesting analogy with the cosmic microwave background (CMB) of the Big Bang universe, which is composed of scattered thermal radiation originating in the earliest phase of the universe that subsequently streamed through the expanding cosmos. From this CMB signature, it is possible to reconstruct the events that occurred right after the birth of the universe. Our findings offer a radically new way to interpret cancer genomic data, providing new insights into how primary human tumors progress, which should facilitate more effective early detection and prognostication efforts.

## METHODS

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

**Accession codes.** The copy number data are accessible via the ArrayExpress database under accession E-MTAB-2140. The sequence data are accessible via the ArrayExpress database under accession E-MTAB-2247. The methylation data are available via the NCBI BioProject database under accession PRJNA230833.

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

### ACKNOWLEDGMENTS

The authors would like to acknowledge the technical assistance of R. Guzman. This project was supported in part by an award to C.C. from the V Foundation for Cancer Research and by award numbers P30CA014089, R21CA149990 and R21CA151139 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the US National Institutes of Health. M.F.P. was supported by a grant from the California Institute for Regenerative Medicine (CIRM).

## AUTHOR CONTRIBUTIONS

A.S., D.S. and C.C. designed the study, interpreted the data and constructed the model. D.S. provided clinical specimens. Z.M. and D.S. processed the specimens. Z.M. generated sequencing data. P.M. and K.S. contributed data. H.K. and M.F.P. performed FISH. A.S. developed and implemented the computational framework. A.S., M.P.S. and J.Z. analyzed the data with oversight from C.C. A.S., D.S. and C.C. wrote the manuscript with input from T.A.G. D.S. and C.C. oversaw the study. All authors read and approved the final 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. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Basanta, D. & Anderson, A.R.A. Exploiting ecological principles to better understand cancer progression and treatment. *Interface Focus* **3**, 20130020 (2013).
4. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
5. Siegmund, K.D. *et al.* Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl. Acad. Sci. USA* **106**, 4828–4833 (2009).
6. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20**, 68–80 (2010).
7. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
8. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA* **110**, 4009–4014 (2013).
9. Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C. & Tavaré, S. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer Res.* **73**, 41–49 (2013).
10. Korolev, K.S., Avlund, M., Hallatschek, O. & Nelson, D.R. Genetic demixing and evolution in linear stepping stone models. *Rev. Mod. Phys.* **82**, 1691–1718 (2010).
11. Korolev, K.S. *et al.* Selective sweeps in growing microbial colonies. *Phys. Biol.* **9**, 026008 (2012).
12. 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).
13. Humphries, A. *et al.* Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proc. Natl. Acad. Sci. USA* **110**, E2490–E2499 (2013).
14. Garcia, S.B., Park, H.S., Novelli, M. & Wright, N.A. Field cancerization, clonality, and epithelial stem cells: the spread of mutated clones in epithelial sheets. *J. Pathol.* **187**, 61–81 (1999).
15. Wright, N.A. & Poulsom, R. Top down or bottom up? Competing management structures in the morphogenesis of colorectal neoplasms. *Gut* **51**, 306–308 (2002).
16. Schwarz, R.F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**, e1003535 (2014).
17. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).
18. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).
19. Park, S.Y., Gönen, M., Kim, H.J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120**, 636–644 (2010).
20. Thirlwell, C. *et al.* Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology* **138**, 1441–1454 (2010).
21. 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).
22. Lopez-Garcia, C., Klein, A.M., Simons, B.D. & Winton, D.J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).
23. Comen, E., Norton, L. & Massagué, J. Clinical implications of cancer self-seeding. *Nat. Rev. Clin. Oncol.* **8**, 369–377 (2011).
24. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
25. Nowak, M.A. *Evolutionary Dynamics* (Harvard University Press, 2006).
26. Lengauer, C., Kinzler, K.W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623–627 (1997).
27. Nowak, M.A. *et al.* The role of chromosomal instability in tumor initiation. *Proc. Natl. Acad. Sci. USA* **99**, 16226–16231 (2002).
28. S Datta, R., Gutteridge, A., Swanton, C., Maley, C.C. & Graham, T.A. Modelling the evolution of genetic instability during tumour progression. *Evol. Appl.* **6**, 20–33 (2013).
29. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep.* **6**, 514–527 (2014).
30. Yatabe, Y., Tavaré, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl. Acad. Sci. USA* **98**, 10839–10844 (2001).
31. Sottoriva, A. & Tavaré, S. in *Proc. COMPSTAT 2010* (eds. Saporta, G. & Lechevallier, Y.) 57–66 (Springer Physica-Verlag HD, 2010).
32. Marjoram, P. & Tavaré, S. Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* **7**, 759–770 (2006).
33. 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).
34. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54–58 (2014).
35. Luebeck, E.G. & Moolgavkar, S.H. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc. Natl. Acad. Sci. USA* **99**, 15095–15100 (2002).
36. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
37. Siegmund, K.D., Marjoram, P., Tavaré, S. & Shibata, D. Many colorectal cancers are 'flat' clonal expansions. *Cell Cycle* **8**, 2187–2193 (2009).
38. Kostadinov, R.L. *et al.* NSAIDs modulate clonal evolution in Barrett's esophagus. *PLoS Genet.* **9**, e1003553 (2013).
39. Burrell, R.A. *et al.* The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
40. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
41. Heng, H.H.Q. *et al.* Stochastic cancer progression driven by non-clonal chromosome aberrations. *J. Cell. Physiol.* **208**, 461–472 (2006).
42. Rosenberg, S.M. Evolving responsively: adaptive mutation. *Nat. Rev. Genet.* **2**, 504–515 (2001).
43. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
44. Sottoriva, A. *et al.* Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity. *Cancer Res.* **70**, 46–56 (2010).
45. Clevers, H. The cancer stem cell: premises, promises and challenges. *Nat. Med.* **17**, 313–319 (2011).
46. Bernards, R. & Weinberg, R.A. Metastasis genes: a progression puzzle. *Nature* **418**, 823 (2002).
47. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
48. Diaz, L.A. Jr. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).

## ONLINE METHODS

**Sample collection.** This study employed deidentified excess tissue specimens collected in the course of routine clinical care and was approved by the local institutional review board (IRB). Individual tumor glands composed of <10,000 adjacent cells were isolated from fresh colectomy specimens following EDTA treatment, as previously described[30]. DNA was isolated from individual glands by incubation in 15 µl of Tris-EDTA solution with Proteinase K solution (4 h at 56 °C), and reactions were boiled for 5 min. Using this method, we consistently obtained samples with >95% tumor purity. Bulk tumor samples and adjacent normal samples composed of a pool of thousands of single glands were also obtained, and DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen).

**Analysis of copy number data.** Individual glands, as well as right and left bulk tumor fragments, were profiled on the OmniExpress SNP platform (Illumina) according to the manufacturer's protocol. Only samples with call rates >85% were analyzed, with an average gland call rate of 97%. Data were processed using GenomeStudio software, followed by quantile normalization[49] and segmentation with psCBS[50], where adjacent normal tissue was employed as a baseline reference for each tumor. To define regions of aberrant copy number, we applied a threshold method based on the standard deviation, $\sigma$, calculated for the 50th central percentile of the probes sorted by the $\log_2$ relative ratio (LRR), adapted from Curtis *et al.*[51]. Briefly, CNAs were determined as follows: amplifications, LRR > 6$\sigma$; gains, 2$\sigma$ < LRR < 6$\sigma$; heterozygous losses, −7$\sigma$ < LRR < −2.5$\sigma$; and homozygous deletions, LRR < −7$\sigma$. The LRR and beta-allele frequency (BAF) for each array were manually inspected to verify the accuracy of the copy number calls and eventually corrected to maintain a conservative approach and to avoid overcalling ITH. Processed copy number data were then used to generate within-gland phylogenetic trees (**Fig. 2b** and **Supplementary Figs. 2** and **4**) using MEDICC[16].

**Analysis of mutational data.** For all adenomas and carcinomas M, N, O, T, U and W, left and right bulk tumor fragments were subjected to whole-exome sequencing to a depth of coverage of 20× on the HiSeq 2000 platform (Illumina). For adenomas K, P and S and carcinomas M and N, the samples subsequently underwent additional sequencing to 60× coverage on the HiSeq 2500 platform (Illumina). For each tumor, a panel of subclonal mutations identified in the bulk fragments and a set of clonal mutations, including putative drivers (for comparison), were profiled in individual tumor glands on the Ion Torrent PGM platform (Life Technologies) using custom AmpliSeq panels. Resultant data were aligned to the hg19 reference genome and processed using MuTect[52] for mutation calling and quantification of allelic frequencies. For the whole-exome sequencing bulk samples, mutations were only called if the coverage exceeded 10× with three or more variant reads. Furthermore, to filter out false positives introduced owing to the presence of paralogous regions, we used BLAST to verify that the 40-bp region around each mutation matched the reference genome uniquely. For the targeted sequencing data, mutations were only called if the coverage exceeded 50× with 20 or more variant reads. To avoid overcalling ITH as a result of false negatives due to low coverage, the absence of a mutation in a gland was indicated not only by a mutation not being called but also by the presence of at least 50× coverage at the locus, of which >95% of the reads had to indicate no mutation. If a mutation was not called in a gland and there was insufficient evidence (owing to low coverage) to confirm its absence, the allelic frequency was annotated as 'NA'. Mutations for which more than half of the glands had NA values were discarded (only four mutations were filtered out because of this problem). The mean coverage for the targeted sequencing data was 626.58 ± 20.2 95% (confidence interval) (**Supplementary Fig. 12**). Public canonical driver mutations (*APC*, *KRAS* or *TP53*) served as a clonal control and are reported alongside the private subclonal events in **Figure 2c** and **Supplementary Figures 1b** and **5**. For tumor O, *APC*, *KRAS* and *TP53* mutations were not detected, and a clonal *CTNND1* mutation is plotted instead. Among the mutations reported, those for which data were available for all glands of a given tumor (where no glands were NA; totaling 167/194 mutations) were employed as input to the statistical inference framework for comparison with the results based on whole-genome copy number profiles. We also employed the mutational profiles of individual glands to infer tumor phylogenies using MEDICC[16]. This allows for the identification of subclones, or groups of glands harboring the same private mutation, where each node in the phylogeny represents a new clone (branching event). By combining the tumor phylogenies and the inferred mutational timelines for different classes of alterations on the basis of our three-dimensional computational model (**Fig. 4** and **Supplementary Fig. 9c**), we could approximately reconstruct the spatiotemporal evolutionary dynamics for each patient (**Fig. 5**).

**Analysis of neutral methylation tag data.** Molecular clock analysis based on neutral methylation tag data was performed as previously described[9]. Briefly, DNA was extracted from individual tumor glands and subjected to bisulfite conversion. Samples were then PCR amplified for the *ZNF454* molecular clock locus, and ultra-deep targeted sequencing (average coverage >1,100× per gland) was performed on the Roche 454/GS JR platform. Data were then processed using our custom pipeline, as previously described[9].

**FISH analysis.** FISH analysis of copy number for the *HER2* gene and chromosome 17 centromere was performed using the Vysis HER-2 DNA Probe kit (Abbott Molecular) in the laboratory of M.F.P., which routinely performs CLIA-certified *HER2* assays. Fluorescence microscopy was employed to quantitatively evaluate the copy number status of 20 cells per gland for 3–6 glands from the left side and 3–6 from the right side of each tumor and 20 cells from 3–4 crypts for each matched normal sample. Thus, 120–240 cells were counted per tumor, and 60 cells were counted per normal sample. Of note, this is 6 times more cells than the 20 that are routinely counted for the diagnosis of *HER2* amplification in breast cancer[53]. As the tissue sections employed for FISH analyses were 5 µm thick, whereas CRC cells are 8–10 µm thick, we verified that this did not introduce bias in estimating the number of amplified cells by analyzing multiple planes and by comparing counts from the tumor and adjacent normal glands (**Supplementary Fig. 13** and **Supplementary Table 3**).

**Computational framework.** We extended our previously described computational framework[9] to (i) accommodate whole-genome copy number and targeted mutational data; (ii) to model fitness effects, corresponding to different survival probabilities; and (iii) to account for microenvironmental niches. This framework exploits ABC, an established approach commonly used in population genetics[32], to obtain posterior parameter distributions by fitting a computational model of tumor growth to the single-gland-level genomic data (**Supplementary Fig. 7a**). The cellular automaton three-dimensional model of tumor growth (**Supplementary Fig. 7b**) accounts for gland growth by fission, the occurrence of CNAs and mutations, and variable gland growth rates. The three-dimensional position of each gland at any point in time is recorded, and glands can have different survival (and growth) fitness owing to CNAs or point mutations. In particular, we simulated the growth of a realistically sized malignancy composed of 8 million glands (~80 billion cells; 5.3 cm in diameter) and incorporated CNAs at a rate $\mu$ that might induce a change in fitness. As simulating changes in fitness for 80 billion cells would be computationally intractable, we assumed that cells within a gland had the same fitness and that fitness changes occurred at the gland level as a result of acquired somatic alterations within the gland (for example, modal copy number changes). Beginning with a single gland with normalized fitness 1 and an associated survival probability, we simulated the possibility that deleterious, neutral and advantageous mutations might change the fitness according to a transition distribution. The input parameters were the mutation rate ($\mu$) and the magnitude of the fitness changes ($\sigma$), where the model produces as output multi-sampling data for each simulated tumor. At the end of the simulation, glands were 'virtually' sampled as they are physically sampled in practice from the tumor, thus maintaining information on the proximity of subclones. In this manner, we faithfully simulated the experimental system (which for practical reasons is restricted to sampling 7–10 glands) several thousand times.

When a CNA occurred, the fitness change was sampled from a Gaussian distribution with mean 0 and variable standard deviation $\sigma$. This models the possibility of both advantageous and disadvantageous alterations. Higher values of $\sigma$ correspond to a greater likelihood that the new clone exhibits an increase or decrease in its fitness, whereas for $\sigma = 0$ no change in fitness occurs, corresponding to the neutral model of growth in which all clones have equal fitness. Here fitness $F$ is expressed in terms of an increase in survival

rate, ranging from 1 to 5, with all simulations initiated with a single gland with $F = 1$. At each division, a gland has probability $P_\alpha = \alpha/F$ of dying, where $\alpha$ is set to 20%. Recent studies indicate the possibility that fitness changes for driver mutations might be as low as 1% (ref. 54), but values on the order of 10% are also typically employed[55]. We evaluated two key parameters, namely, the magnitude of fitness changes $\sigma \in \{0, 0.2, 0.6\}$, corresponding to no change, moderate and large changes, respectively, and the mutation rate $\mu \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$ per gland per division. Because $\sigma$ is the standard deviation of a normal distribution with mean = 0, for $\sigma = 0.2$ we expect a fitness increase of 10% or greater in 30.7% of the cases, whereas for $\sigma = 0.6$ a fitness increase of 10% or greater is expected in 48.8% of the cases. Thus, these values correspond to a range of small to large variations in fitness. Other complex and poorly characterized processes, such as cellular migration and apoptosis within a gland, are not modeled, nor is the contribution of the surrounding normal tissue or angiogenic factors.

The simulation began with a gland in the center of a $400 \times 400 \times 400$ point lattice where glands then split by fission until a volume of 8 million glands is reached. Subsequently, five glands from the left side and five glands from the right side of the tumor were virtually sampled from the simulation, in accordance with the experimental sampling scheme performed on the tumor specimen. The CNA profiles of the sampled glands were saved for comparison with the actual data (**Supplementary Fig. 7a**). We employed ABC to fit the model to the data, to generate posterior probability distributions of the parameters ($\sigma$ and $\mu$) for each patient, assuming uninformative uniform priors. Every CNA was associated to a binary string indicating its presence (1) or absence (0) in each sampled gland. Public alterations were excluded from the inference, as the vast majority likely occurred during preneoplastic stages, before the transition to an established neoplasm, and thus do not belong within the simulated scenario. Nevertheless, relaxing this rule yielded similar results (data not shown). Summary statistics were then computed using these binary patterns, including the number of distinct CNAs (the number of different strings), the Shannon index of the binary patterns, the total number of alterations, the number of variegated alterations and the number of side-variegated alterations. As a measure of the distance between the actual data and the simulated data, we employed the average distance of the summary statistics, normalized to mean = 0 and s.d. = 1. The inference framework was validated using synthetic data to demonstrate that the correct parameter value was accurately recovered in the majority of cases (**Supplementary Fig. 14**).

To examine the influence of differences in local tumor microenvironment, we developed a version of the model in which specific CNAs were selected depending on the surrounding tumor area by incorporating static microenvironmental niches of differing size (env parameter: $5 \times 5$, $20 \times 20$, and $150 \times 150$) in the simulation. Each 'block' in the grid selects for a random CNA or mutation on a specific chromosome by inducing a high apoptosis rate (20%) for glands that do not have that particular alteration, such that the overall apoptosis rate is quite high, representing positive selection. In this manner, rudimentary microenvironmental niches that select for different gland populations are represented (**Supplementary Fig. 10**). The same approach as described above was applied to perform inference on mutational profiles in both the niche-based and microenvironment-free models (**Supplementary Fig. 11**). The results imply that distinct yet static local microenvironments do not alter Big Bang dynamics. In the future, it will be of interest to examine contributions due to dynamic interactions between tumor cells and their microenvironment, as well as clonal cooperation and interference.


49. Staaf, J. *et al.* Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics* **9**, 409 (2008).
50. Olshen, A.B. *et al.* Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27**, 2038–2046 (2011).
51. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
52. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
53. Wolff, A.C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* **25**, 118–145 (2007).
54. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA* **107**, 18545–18550 (2010).
55. Michor, F., Iwasa, Y. & Nowak, M.A. Dynamics of cancer progression. *Nat. Rev. Cancer* **4**, 197–205 (2004).





doi:10.1038/ng.3214