Evolving copy number gains promote tumor expansion and bolster mutational diversification

Zicheng Wang[1,2†], Yunong Xia[1,2], Lauren Mills[3], Athanasios N. Nikolakopoulos[1,2], Nicole Maeser[1,2], Jason M. Sheltzer[4] and Ruping Sun[1,2*†]

[1*]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, Minnesota, United States of America.
[2]Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota, United States of America.
[3]Department of Pediatrics, University of Minnesota, Minneapolis, Minnesota, United States of America.
[4]School of Medicine, Yale University, New Haven, Connecticut, United States of America.

*Corresponding author(s). E-mail(s): ruping@umn.edu;
Contributing authors: wang2569@umn.edu; xia00045@umn.edu; ljmills@umn.edu; annikolako@gmail.com; maese004@umn.edu; jason.sheltzer@yale.edu;
†These authors contributed equally to this work.

Abstract

Charting the evolutionary history of rampant somatic copy number alterations (SCNA) is an indispensable step toward a deeper understanding of their roles in tumor development. However, the existing SCNA timing analysis is limited to low copy number states and initiating gains, which are not necessarily close to the onset of the malignant proliferation. Moreover, it remains a critical question if the timing of an SCNA reveals the corresponding variant's fitness effect. Here we propose a framework to estimate the arrival time of a clonal SCNA, i.e., the time delay from its last alteration to the start of population expansion, in addition to its initiation time when the first alteration occurs. Our method leverages the bias

that a genomic segment, when resting on a copy number (CN) state, accumulates somatic single nucleotide variants (SSNV) at a rate proportional to its CN. From the whole genome sequencing data of 184 samples from 75 patients across five tumor types, we commonly observed late clonal CN gains following early initiating events, which appear right before the expansion leading to the observed tumor sample(s). Genome doubling (GD) can be late, but post-GD CN evolution is prevalent in the genealogy of the most recent common ancestor of patient tumors. Notably, mathematical modeling suggests that late evolving events could contain rate-limiting drivers. The advantage of evolving gains could arise from the dosage increase of cancer genes in proliferative signaling and amplification of early functional variants. In addition, evolving SCNAs bolster the diversification of SSNVs between sub-populations, exacerbating the vicious circle between malignant growth and accumulation of genomic heterogeneity. Our findings have broad implications for resolving the trajectory of SCNAs, discerning the CN markers of malignant growth *in vivo*, and properly quantifying tumor heterogeneity in aneuploid tumors.

Keywords: Timing of Somatic Copy Number Alteration, Aneuploidy, Whole genome sequencing, Cancer evolution, Tumor heterogeneity

1 Introduction

Underlying the maintained genomic diversity within a patient tumor is the uncontrolled proliferation, a core hallmark of cancer [1], coupled with somatic alterations occurring over time [2]. To prevent the disease, uncovering the somatic aberrations responsible for the malignant growth is the primary goal of precision oncology. At the genomic level, somatic alterations exist on a spectrum, ranging from small changes such as somatic single nucleotide variants (SSNV, often referred to as mutations) [3] to large somatic copy number alterations (SCNA). Frequent chromosomal mis-segregation (chromosomal instability or CIN) leads to abnormal chromosome numbers (aneuploidy[4]) and unbalanced structural variations (SV) cause segmental SCNAs [5]. These two genomic errors are intertwined in many solid tumors, leading to extensive SCNAs, especially in advanced diseases [4] with poor clinical outcomes [6].

The inextricable relation of SCNAs to cancer initiation[7, 8] and progression[9] has become a consensus in cancer genomics. It remains little known, however, to what extent a specific SCNA accounts for the malignant growth and how it affects the intra-tumor-heterogeneity (ITH) [4]. Indeed, chaotic karyotype and widespread high copy number (CN) states in aneuploid tumors[10] make it extremely challenging to identify SCNA drivers, limiting the precision of using SCNA patterns for diagnostic and treatment purposes. For example, the treatment strategy for osteosarcoma, the most common bone tumor affecting teenagers with one of the most chaotic aneuploid genomes, has stagnated for decades [11]. From an evolutionary perspective, discovering

the tempo of SCNA during somatic evolution is key to gaining knowledge of SCNA drivers [12]. Here, we hypothesize that the timing of SCNAs can be systematically measured from whole genome sequencing (WGS) data of patient tumors, and the temporal axis contains tangible information in isolating the effect of specific SCNAs on tumorigenesis.

We should pause to clarify how bulk sequencing data capture somatic evolution timeline. The tumor founder cell arose from the succession of clonal expansions in the pre-cancerous context where beneficial alterations endow progenitor cells with the ability to crowd out less advantageous populations [13] (Figure 1A). The growth of the primary lesion gives rise to genomically diverging lineages [14], some of which, after acquiring a more malignant potential, can initiate the re-growth of a secondary tumor, such as metastasis [15]. Bulk sequencing data provide us with the opportunity to anchor the roots of expansion (the most recent common ancestor, or MRCA). Clonal variants in a single sample refer to the root of the observed sample. In multi-region sequencing, truncal variants from multiple samples could trace back asymptotically to the founder of the tumor [16]. For longitudinal sampling, e.g., of paired primary and metastatic tumors, truncal variants could point to the MRCA of the branched tumor progression [17]. Multi-samples reflect the population expansion at a broader scale, i.e., they coalesce to an earlier progenitor cell than a single sample does. Collectively, truncal variants revealed by a particular sampling strategy must map to the somatic evolution timeline prior to the corresponding *sampling-relevant expansion* (SRE).

The timing of a truncal SCNA (more specifically, gains) on the evolution toward the MRCA could shed light on the impact of this variant on promoting the SRE. However, our knowledge about the SCNA timing remains fragmentary as the existing methods are restricted to simple (single or double) gains [16, 18, 19]. Single-sample based pan-cancer analysis from ICGC[19] revealed extensive variation in the molecular time of these sample gains. Relative to aneuploid tumors, these gains are likely the initiating events of copy number evolution and may not be sufficient to induce the expansion. For example, single chromosome gains have shown limited capability in driving proliferation in *in vitro* systems [20]. As such, it would be critical to know when a genomic segment further evolves beyond the initial gains, which can often culminate at a state greater than four copies [10]. This requires a timing method applicable to high SCNA states. In SCNA timing analysis, the following assumption is made: the site frequency spectrum (SFS) of SSNVs in a genomic region affected by SCNA depends on the trajectory (the order) and time span on each CN state that the segment had ever rested on [21]. For a single gain, the ratio between early (duplicated) and late (non-duplicated) SSNVs can be used to estimate the relative timing of the gain [12]. For a high CN state (> 4) or gains involving both maternal and paternal alleles, the detailed trajectory of each SCNA evolution is not identifiable from the SFS [18], at least not without further constraints on the trajectory or additional information that could

allow one to reconstruct the ordering [21]. We reason, however, that it is possible to obtain the upper bounds of the time duration of a segment resting on its observed truncal CN state before the onset of SRE. In principle, once a segment arrives at the final CN state, it can only accumulate single-copy SSNVs; the longer the segment persists in the observed CN state, the more overwhelming the single-copy SSNVs. Notably, for the first time, this allows us to investigate the *arrival time* for SCNAs at high CN states.

An outstanding question is how the timing of an SCNA reflects its fitness effect (Figure 1B). Whereas early gains could initiate and increase the risk of disease, we argue that late-appearing SCNAs close to MRCA could promote the population expansion more directly. If an SCNA triggers SRE, one may foresee that the progenitor cell proliferates vigorously upon acquiring the corresponding SCNA. By contrast, before the SRE begins, if the clonal lineage persists for many generations after receiving an SCNA, the corresponding variant is unlikely to be sufficient to drive the SRE. Punctuated acquisition of polyploidy (e.g., through genome doubling or GD) is prevalent in aneuploid tumors [22] but it remains unclear how close the occurrence of GD is to tumor transformation. Evidence exists that GD itself doesn't confer a strong fitness advantage[23]; instead, it can enhance the plasticity of the genome that permits further CN evolution, such that aneuploid cells adapt to overcome possible fitness penalties incurred by GD [24]. Therefore, SCNAs that arrive after GD could contain driver events. Moreover, depending on the precise location of biopsied tissue, single-sample analyses may differ in the corresponding time scale; subsequently, it is particularly essential to focus on the timeline toward the malignant growth - the somatic evolution in collecting truncal SSNVs of multiple samples of a tumor (e.g., multi-region samples or paired primary and metastatic samples).

To broaden the "timable" genomic regions for SCNAs, we developed `Butte` (BoUnds of Time Till Expansion), a computational framework to estimate the upper bounds of SCNA arrival and initiation time from WGS data. By applying `Butte` onto multi-sampled WGS data of five cancer types with widespread SCNAs, including osteosarcoma, we systematically charted the temporal patterns of CN evolution *in vivo*. To see if late-appearing SCNAs may confer fitness benefits, we constructed mathematical models to examine the evolutionary mechanisms that give rise to these late truncal events. Furthermore, we also interrogated potential ways the late culminating SCNAs could add to the fitness and revealed its impact on mutational diversification during tumor expansion.

2 Results

2.1 A computational framework to estimate the arrival time of SCNA gains

From the WGS data, one can characterize with high certainty the dominant SCNAs, inferring the integer allelic CN of a genomic region and the cellular

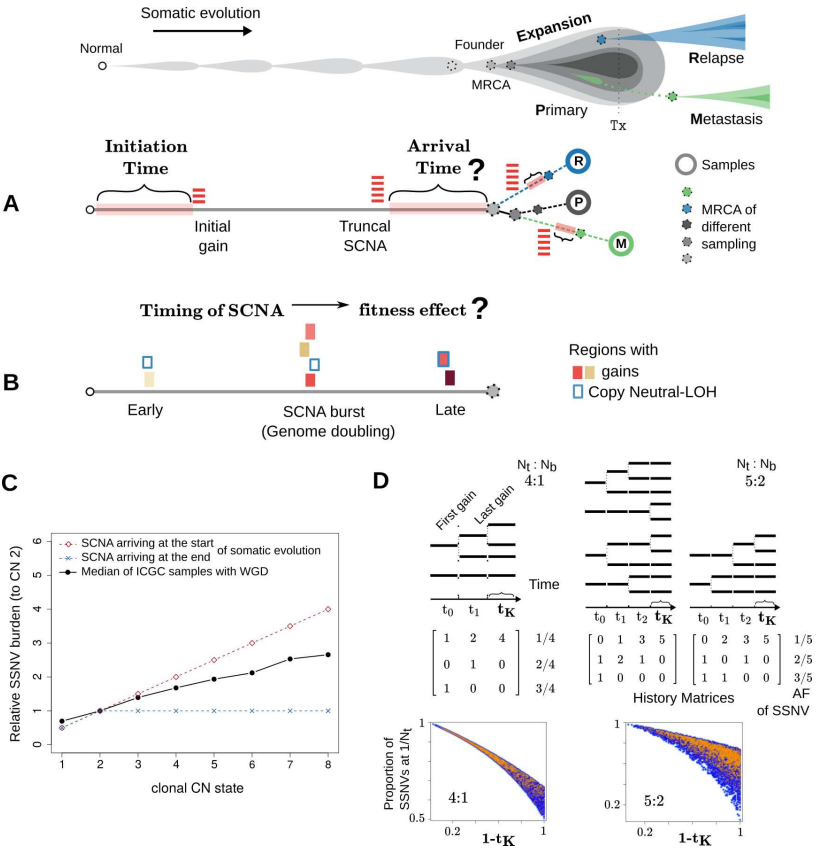

Fig. 1 Measuring the arrival and initiation time of SCNAs. A drawing at the top marks the concept of cancer somatic evolution, which leads to the establishment of the most recent common ancestor (MRCA) of the primary tumor, as well as the roots of secondary tumors. Tx: treatment. (A) For an SCNA present within the MRCA of a sampling-relevant tumor expansion, we aim to characterize the time when the last gain and first gain appeared (referred to as arrival and initiation time, respectively) for the corresponding genomic region. (B) We further aim to address the question if the late truncal gains are neutral or beneficial to tumor evolution. (C) The solid black line shows that the burden of SSNVs of a given genomic locus correlates with its CN state in tumors sequenced by ICGC (single sampling, with WGD). Two dashed lines assume the two extreme scenarios of SCNA arrival time. (D) The proportion of SSNVs at different allele fractions depends on the SCNA history matrix and the relative time span of each CN stage. The smoothed scatter plots show the burden of single-copy SSNVs against the arrival time of an SCNA (t_K), with an increasing density from blue to orange. Two example SCNA states are shown (with CN configuration at 4:1 and 5:2, respectively).

prevalence of the corresponding SCNA, i.e., the percentage of cells sharing the dominant SCNA state [25]. We refer to a unique version of a genomic region (or segment) as an "allele". We term the total CN as N_t and the CN for the minor allele as N_b ("b" stands for b-allele determined by germline SNPs) for a dominant SCNA. The "allele" fraction of an SSNV is the amount out of the total N_t copies of the region that carry the corresponding variant. We

found that in the aneuploid tumors sequenced by ICGC (International Cancer Genome Consortium), the SSNV burden increases with the dominant SCNA states of the corresponding genomic region (Figure 1C). Such a pattern is due to an intrinsic bias between SCNAs and SSNVs: a genomic segment resting on a CN state accumulates SSNVs at a rate proportional to the corresponding CN. Thereby, the burden and CN multiplicity of SSNVs are actively shaped by SCNAs.

Generally, the observed SCNA of a genomic segment (with a configuration $N_t : N_b$ different from $2 : 1$) is the result of a series of CN events. For an SCNA involving at least K gain events, the total time of somatic evolution can be divided into $K + 1$ stages. The genomic segment begins with the $2 : 1$ setting in the first stage and keeps "climbing up" by duplicating one of its existing copies in each subsequent stage, respectively, until it arrives at the observed SCNA state in the last stage (Figure 1D). Accordingly, each stage is associated with certain time proportion ($t_k \geq 0$) and $\sum_0^K t_k = 1$, the total time for the somatic evolution. SSNVs occurring at stage k on a segment copy that experiences duplication(s) in later stages will be presented on the duplicated copies ($a \geq 2/N_t$). By contrast, SSNVs acquired at stage k on a copy without further duplication events remain at the single allele state ($a = 1/N_t$). One can define a history matrix A with entry A_{jk} representing the number of segment copies in stage k that result in a final allele state a_j [18, 21]. It can be seen that the abundance of SSNVs at allele state a_j depends on $\sum A_{jk}t_k$. From the site frequency spectrum (SFS) of SSNVs in a region affected by SCNA, one can estimate the relative abundance of SSNVs at each allele state, and in turn, solve for each t_k. There has been much effort to infer t_0, i.e., the timing of the first copy number event [18, 19]. These efforts focused on single gain ($2 : 0$ and $3 : 1$) and at most double gains ($3 : 0$ and $4 : 1$), where the history matrix A is invertible. By contrast, for other SCNA states, multiple possible trajectories can exist and the underlying linear system is underdetermined, i.e., there are more time stages (unknown variables) than the possible allele states (equations). We note that, however, regardless of the underlying history, multi-allele SSNVs ($\geq 2/N_t$) can only occur before the last stage (K) of CN evolution; once the genomic region arrives at the observed clonal SCNA state, all the copies (N_t) would accumulate SSNVs at single allele state ($1/N_t$). Therefore, the longer the last stage of CN evolution (from the emergence of the clonal SCNA to the onset of population expansion), the more overwhelmingly the single allele SSNVs dominate the SFS (Figure 1D). Such monotonicity allows one to calculate the bounds of the time duration for the last stage even when the detailed SCNA history is unidentifiable.

To investigate how various SCNAs unfold during somatic evolution, we developed `Butte` (BoUnds of Time Till Expansion), which adopts linear programming to infer the upper bounds of arrival (t_K) and initiation time (t_0) of SCNAs (see Methods). From a technical point of view `Butte` extends the full

maximum-likelihood estimation procedure implemented in `cancerTiming` [18]. Notably, `Butte` does not restrict the analysis to single and double gains, but in addition allows the calculation of the upper bounds of t_K and t_0 for SCNAs up to seven total copies, broadening the "timable" SCNA regions. The upper bounds, by definition, overestimate the time duration. Nevertheless, these estimates systematically correlate with the actual timing of SCNA initiation and culmination (Supplementary Figure 1). To test the performance of `Butte` on real patient tumors, we first evaluated the timing predictions by analyzing multi-region WGS data of colorectal adenocarcinomas (COAD) [16, 26]. `Butte` successfully identified early CN gain of chromosome (chr) 5q (Supplementary Figure 2), corresponding to the SCNA state of $2:0$ (copy neutral loss of heterozygosity), a known early step in COAD initiation involving gene *APC* [27]. As a benchmark for late-appearing events, private (sample-specific) SCNAs should contain events that occur in the descendent lineage of the MRCA of multi-samples. `Butte` predicted their arrival time to be later than the public SCNAs on the timeline toward the MRCA, highlighting its capability in revealing the late arriving SCNA events (Supplementary Figure 3).

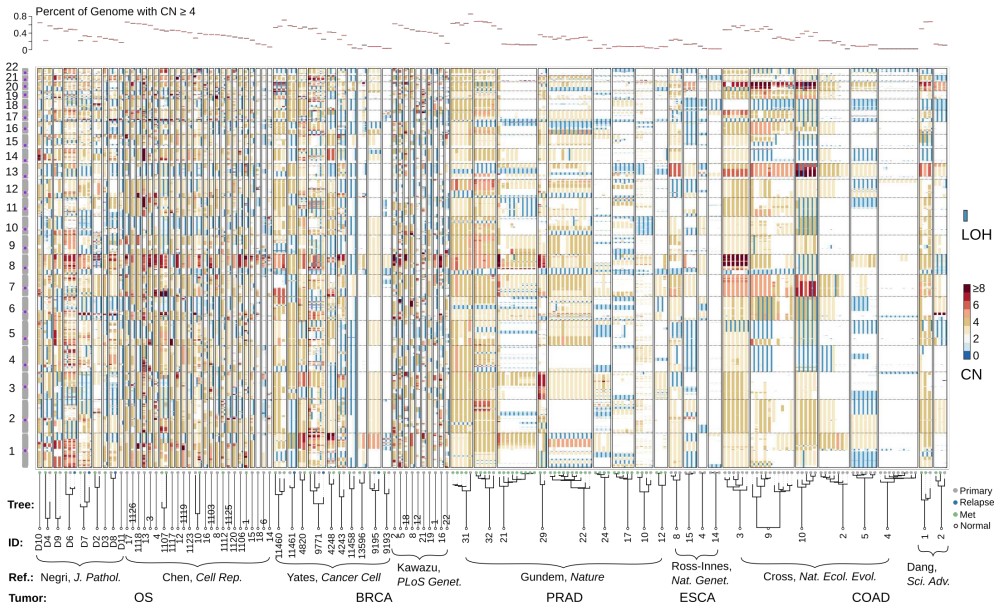

Fig. 2 CN profiles across five tumor types from the re-analysis of published WGS data. Vertical bars represent the segmental CN states along the autosomal chromosomes characterized by the WGS data of tumor samples. For each sample, a color-coded thick bar shows the total CN state of each genomic locus and a thin blue bar to the right indicates that region has loss of heterozygosity (LOH). Samples belonging to the same patient are boxed. The top panel highlights the fraction of high CN states in each sample's genome. The lower panel exhibits the sample phylogenetic trees constructed from SSNVs. Sample IDs, the reference where the WGS data was published, and tumor types are tabulated at the bottom. In this manuscript, a tumor sample is named after the concatenation of the tumor type, the first character of the author's surname and the patient ID.

2.2 Evolving SCNA gains define the tumor transformation leading to the most recent clonal expansion

To evaluate the tempo of SCNAs in solid tumors, we applied `Butte` on five tumor types by analyzing eight published WGS datasets: osteosarcoma (OS) [28, 29], breast invasive carcinoma (BRCA) [30, 31], colorectal adenocarcinomas (COAD) [16, 26], esophageal carcinoma (ESCA) [32], and prostate adenocarcinoma (PRAD) [33], six of which comprise multi-sampling of patient tumors (Figure 2, Table 1). 70% of the analyzed genomes (corresponding to 87% of the patients) were near triploid, with the median fraction (IQR) of the high amplitude CN regions (≥ 4) being 0.37 (0.23 to 0.49). Loss of heterozygosity (LOH) is prevalent but mostly is at copy neutral or amplified states in the triploid tumors. High amplitude gains can be recurrent across cancer types (e.g., chr 8q) or within a specific tumor type (e.g., chr 1q for BRCA, chr 17p for OS, and chr 7 for COAD). These recurrent gains presumably contain driver events [34], yet their tempo in somatic evolution remains uncharted. Notably, karyotypes largely remain stable across different samples of the same tumor, despite the presence of continued subclonal CN diversification in a relatively minor fraction of the genome.

We note that 74 out 75 patient tumors acquired late-appearing gains close to MRCA regardless of the overall ploidy or tumor type (Figure 3A, B), with the only exception of COAD_C_4, which shows high microsatellite instability. Punctuated copy number bursts were observed in the triploid samples, reflecting the ability of the genome to leapfrog over intermediate states to reach moderately high CN states through whole or partial genome doubling (GD)[35, 36]. Whereas GD occurs late (close to the MRCA) in some adult cancers (18 out of 34 patients), it appears to be an earlier event in many other tumors. This is particularly evident in OS where 28 of 30 patients had GD at the mid-stage of somatic evolution toward SRE (Figure 3B). The contrasting tempo of GD suggests that it probably has a context-dependent fitness. In tumors with early GD, `Butte` can characterize the post-GD CN evolution, whereby progenitor cells continue to sample the aneuploid fitness landscape [24]. Such an SCNA evolution involves stochastic chromosomal or structural abnormalities; however, certain genomic regions preferentially exhibit late gains across different patients in a particular tumor type, which, surprisingly includes those recurrent high amplitude gains, such as chr 8q in OS (Figure 3C) and chr 7 in COAD (Supplementary Figure 2). On the other hand, recurrent SCNAs appear to initiate early, e.g., chr 1q (*AKT3*) in BRCA, chr 8q and chr 17p (*TP53*) in OS and chr 5q (*APC*) in COAD (Figure 3C and Supplementary Figure 2). These additional gains pre- and/or post-GD could result from either high evolvability of the corresponding region, or persistent selection upon driver genes within.

The earlier the timing of GD, the more post-GD CN gains (Figure 4A). The late evolving gains are shorter in segment length than those associated with GD (Supplementary Figure 4), except for a few patients where post-GD events

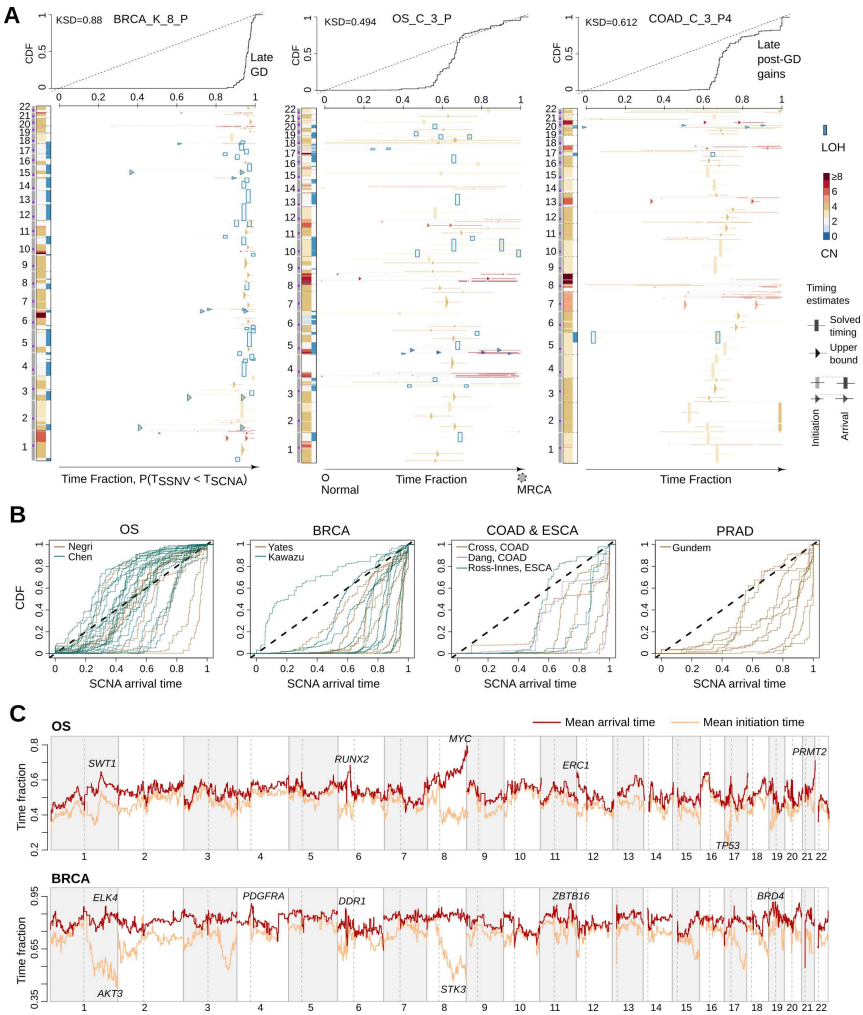

Fig. 3 The timing patterns of SCNAs across five tumor types. (A) Timing plots of three exemplified tumors. CN states along the genome are shown on the left of each panel. The right panel visualizes the time fraction of somatic evolution from germline to the MRCA of the patient tumor. For each SCNA segment, the inferred timing is drawn by either rectangle (exactly solved timing) or arrows (upper bounds of timing when the solutions are not unique) with the same color-coding as its CN. The top panel shows the cumulative distribution (CDF) of SCNA arrival time. (B) The CDF curve of SCNA arrival time is shown for each patient categorised by the tumor type. (C) The average timing (both initiation and arrival time) for each one Megabase genomic bin across OS and BRCA patients are shown, respectively. Example genes in regions showing late arrival or early initiation times are marked.

are sparse (e.g. for patient PRAD_G_21 and PRAD_G_31). This suggests that the post-GD CN evolution is driven by SVs, which occur at a higher rate than chromosomal mis-segregation. Indeed, the breakpoints of structural variants almost locate the boundary of SCNA segments (Supplementary Figure 5). As

SVs continued to occur, it could become more focally amplified for the driver genes that are repetitively selected, making them more apparent in the late appearing gains, e.g., *MYC* [37] and *RUNX2*[38] in OS (Figure 3C). In terms of types of gains, amplified LOH ($N_b = 0$ and $N_t >= 3$) tend to culminate later than other types of amplifications, such as allele specific amplifications ($N_b = 1$ and $N_t >= 3$, $p < 3e-5$), and this cannot be explained by the overestimation of `Butte` (Supplementary Figure 6 and 1). Whereas truncal LOH were supposedly acquired before GD [15] causing the complete loss of tumor-suppressor activity, the late appearing gains of the only remaining allele may indicate that these regions potentially acquire dosage-dependent gain-of-functions [39].

2.3 Mathematical modeling suggests the role of late gains in promoting tumor expansion

While early genomic changes during somatic evolution are suggested to be functional initiating events, there is little known about the effect of late-appearing alterations (e.g., close to the MRCA). To reveal the mechanism that gives rise to the tempo of GD and the number of post-GD events, we mathematically modeled the somatic evolution based on a multi-type branching process (see for example [40]), starting with a single tumor-initiating cell that just acquired GD (Figure 4B). This cell gives birth at a rate of a_0 and dies at a rate of b_0 (with net growth rate $\lambda_0 = a_0 - b_0 > 0$). During progression, daughter cells acquire a passenger post-GD gain with a rate of u_0, which does not change the net growth rate. With a lower rate u_1, they can also acquire a driver post-GD gain, leading to increased fitness ($\lambda_1 = a_1 - b_1 > \lambda_0$, see Method for details). Our objective is to characterize the number of post-GD gains that reach fixation or become dominant in the observed tumor under two contrasting scenarios: one without and the other with driver post-GD gains.

We first assumed that all post-GD gains are passenger variants. Conditioned on the non-extinction of the population, we obtained that the number of post-GD gains accumulated before the MRCA that grows into detectable tumor follows a geometric distribution with parameter $\frac{\lambda_0}{\lambda_0+u_0}$ and mean $\frac{u_0}{\lambda_0}$. The mode of this distribution is at zero, similar to the cases where GD appears late and post-GD CN gains are rare. To tolerate the inclusion of subclonal but dominant SCNAs as the clonal variants, we further evaluated the dominant post-GD gains shared by the majority ($\geq 90\%$) of cancer cells. Building on the results of [41], we derived the expected number of dominant post-GD gains in a tumor with size N as

$$\tilde{S} = \frac{N}{\lceil 0.9N \rceil} \cdot \frac{u_0}{\lambda_0} \approx 1.11 \frac{u_0}{\lambda_0}, \tag{1}$$

which is only slightly larger than the clonal ones. Assuming that u_0 and λ_0 are comparable, $\tilde{S}$ would be no more than just a few. Moreover, numerical

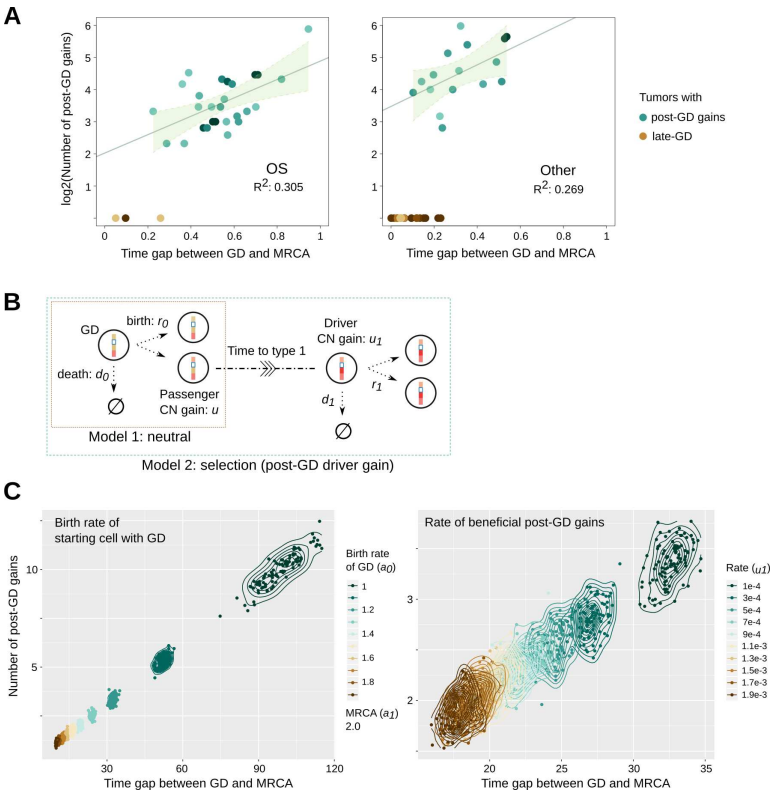

Fig. 4 Mathematical modeling suggests that advantageous CN gains can occur late. (A) Scatter plots colored by density illustrate the number of post GD gains against the time fraction of post-GD evolution towards MRCA for OS and other adult cancer types, respectively. A linear regression line and the confidence intervals are indicated for tumors with post-GD gains. (B) The schema shows the setup of the two contrasting mathematical models: (1) GD is followed by neutral growth where additional gains do not confer a fitness advantage and (2) post-GD gains increase the growth rate. (C) Scatter plots display the number of post-GD passenger gains against the time of post-GD evolution characterized by the selection model. We studied the effects of the growth rate of GD (with a fixed growth rate of the MRCA, the left panel) and the rate of beneficial post-GD driver gains (the right panel), respectively. Each point refers to the average of 100 simulations. Contours of the 2d density estimation are shown. Parameters for the left panel: $b_0 = 1$, $a_1 = 2$, $b_1 = 1$, $u_0 = 0.1$, and $u_1 = 0.0001$; right panel: $a_0 = 1.2$, $b_0 = 1$, $a_1 = 2$, $b_1 = 1$, and $u_0 = 0.1$.

simulations show that the number of dominant post-GD gains continues to follow a geometrical-like distribution with the mode at zero. Thus, if post-GD gains do not provide growth benefits, GD would be one of the last events before the MRCA as few of post-GD gains can become dominant in the observed tumor.

However, if post-GD gains increase fitness, the situation drastically changes. To emphasize how this happens, let us consider the first advantageous cell (type 1), which is introduced via a beneficial post-GD gain and grows into an infinite number of descendants. The original population without the new

driver forms the type 0 population. In this scenario it suffices to focus on type 1 cells as their population most likely overshadows the type 0 population upon detection. The expected number of passenger post-GD gains ($\bar{S}$) carried by a type 1 cell would be proportional to the time of occurrence of the type 1 cell. In Methods we show that the distribution of the birth time of the first non-extinct type 1 cell, $\mathbb{P}(\sigma_1 > t \mid \Omega_\infty)$, where σ_1 represents the birth time and Ω_∞ represents the event that the population does not go extinct, can be characterized as a function of the rate of beneficial gains u_1 and growth parameters of type 1 and type 0 cells, respectively (Lemma 1). $\bar{S}$ is thus,

$$\bar{S} = \int_0^\infty \mathbb{P}(\sigma_1 > t \mid \Omega_\infty)\, u_0 dt. \tag{2}$$

We explored various choices of growth parameters that capture the fitness difference between type 0 (with GD) and type 1 cells (with advantageous post-GD gain). Notably, lowering the fitness level of type 0 cells delays the birth of the type 1 cell (Figure 4C), conditioned on a fixed net growth rate of the type 1 cell. Accordingly, the post-GD gains become abundant with a bell-shaped distribution when the fitness difference is large. These model predictions are consistent with timing patterns in tumors with early GD (Figure 4A). Therefore, early GD observed in many patient tumors suggests that late-appearing gains may confer additional advantages for promoting the expansion. We note that the prolonged period of post-GD evolution could also be attributable to a lower rate of beneficial post-GD gains (Figure 4C). Nevertheless, the beneficial gain inevitably appears to be a late event since passenger gains introduced after the beneficial one most likely will be undetectable. The inability of passenger variants to become dominant themselves ensures such an outcome. Therefore, a long waiting time after GD suggests that the late beneficial gain becomes a rate-limiting event.

2.4 Ways evolving CN gains contribute to fitness increase and mutational diversification

As SCNAs have a global impact on gene expression in cancer [42], the evolving CN gains potentially affect dosage-sensitive genes whose gains have a functional impact. In the OS and BRCA tumors, as the CN evolves, we can see an enrichment of putative dosage-sensitive genes that are in pathogenic CNV peak regions derived from dbVar [43, 44] (Figure 5A). Notably, we observed a similar enrichment for genes involved in sustaining proliferative signaling: one of the most fundamental capabilities of cancer cells [1]. *MYC*, *EGFR* and *KIT* are among such genes with late gains in both OS and BRCA, emphasizing their ability in stimulating cell multiplication in multiple tumor types.

The evolving gains could amplify the impact of early functional variants by increasing their multiplicity (Figure 5B). Such a mechanism potentially affects SV breakpoints in known oncogenes (e.g., *MAP3K13*, *MECOM* and

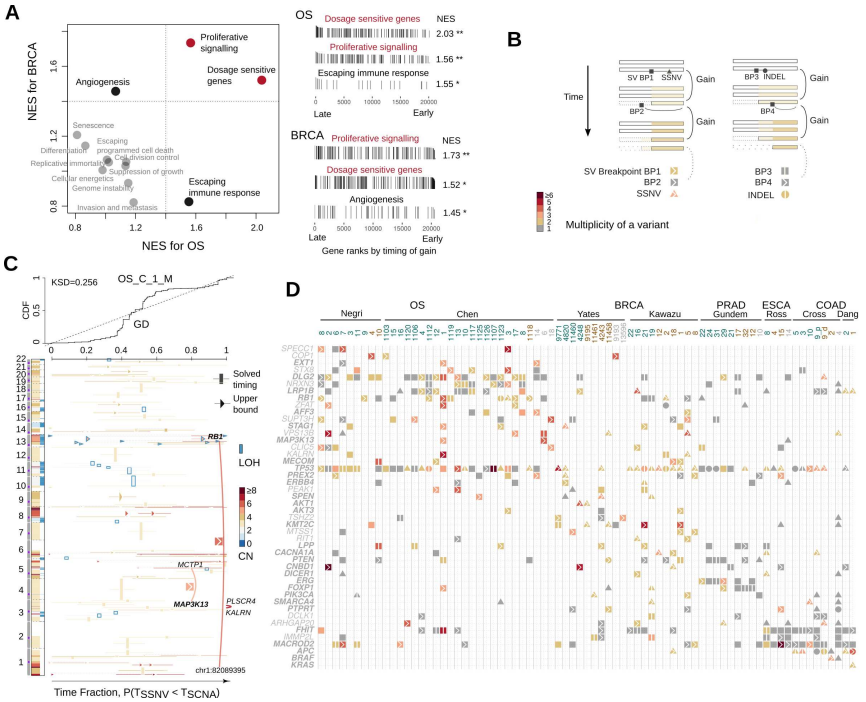

Fig. 5 Ways the late CN gains contribute to the fitness of cancer progenitor cell. (A) The gene set enrichment analysis (GSEA) was performed on the gene list ranked by the averaged CN arrival time for BRCA and OS tumors, respectively. The scatter plot on the left shows the normalized enrichment scores (NES) for each set of cancer census genes belonging to the predefined cancer hallmarks by COSMIC database. The vertical bars on the right panels visualize the timing-ranks of genes that belong to the highlighted gene sets. (B) A cartoon illustrates the multiplicity increase of an early sequence variant due to the inclusion of that variant by a late CN gain, with annotations indicating the type of variants (symbol shapes), level of multiplicity (color hues) and the variants' association with a late gain (right arrow) or an early gain (vertical bar). (C) The SCNA timing plot of an example OS tumor similarly arranged as in Figure 3, with additional links and symbols highlighting the SV breakpoints in known cancer genes that are amplified by late gains. (D) The matrix plot demonstrates genes with recurrent somatic variants and their multiplicity across the five tumor types. Names for known cancer genes are in bold. Genes with variants showing higher multiplicity levels than gene TTN are also included. Symbol annotations are the same as in (B).

PREX2), breakpoints in genes known to be involved in oncogenic fusions (e.g., *AFF3*, *LPP* and *ERG*), and simple mutations in oncogenes (e.g., SSNVs in *SMARCA4* and *CACNA1A*), see Figure 5C, D. *MAP3K13* had been shown to promote tumor growth in high *MYC*-expressing cells [45, 46], a similar context as in the OS [37]. We note that highly mutated tumor suppressor genes (TSG), such as *TP53*, *RB1* and *APC*, also have their early mutants duplicated or amplified (Figure 5D). Whereas these are presumably inactivation variants, the retaining of multiple copies of the variants could suggest different roles that remain unclear, such as a potential gain-of-function of APC mutants in COAD [47]. The fact that SRE requires the duplication of these early variants,

rather than starting immediately upon acquiring a single copy of these variants, suggests that late-appearing gains could cooperate with the early variants to promote tumor expansion. On the other hand, late SV breakpoints (at single copy state) coupling evolving gains are prominent in genes located in common fragile sites, e.g., *FHIT* and *MACROD2*. Late alterations of these genome "caretakers" could facilitate further genome evolution and expedite clonal expansion [48, 49].

Lastly, the quantitative relation between SCNA evolution and SSNV accumulation, the rationale behind our timing method, implies that SCNA gains bolster mutational diversification between sub-populations during tumor growth. In principle, the higher the truncal CN state of a genomic segment, the higher the mutational divergence between subclones for the corresponding locus. As tumor expands, genomic regions at distinct SCNA states would accumulate SSNVs at different rates, leading to the heterogeneity of the SSNV burden along the genome. For example, when comparing two samples of a tumor, the sample-specific SSNVs are more abundant for regions with higher CN states (Figure 6A, B). Notably, the overall CN state affects the structure of phylogenetic trees, i.e., it explains more than 50% of the variance of the relative branching distance measured by SSNVs in COAD and PRAD patients, where extensive multi-region sampling is available (Figure 6C). Furthermore, continued evolution of SCNAs between subpopulations would also alter the SSNV divergence. For example, the SSNVs divergence is particularly enlarged for regions showing different CN states between the two samples (Figure 6B). As such, increased SSNV diversity in regions with CN gains provides more materials for further selection within the expanding cell populations.

3 Discussion

The major impediment to devising better therapies for tumors with the CIN phenotype is its highly complex somatic genome. In this study, we have created a computational framework for measuring the arrival and initiation time of SCNAs during the somatic evolution of the MRCA of tumor sample(s). By applying this method on WGS data of patient tumors, we have found that late truncal CN gains close to the most recent clonal expansion are common across multiple tumor types. Mathematical modeling predicts that these late evolving gains could contain rate-limiting driver events, promoting the tumor growth. As CN gains increase the gene dosage and early functional variants, we further demonstrated that an integrated analysis of SCNA timing, SV breakpoints and simple mutations has a strong potential for isolating the functional effect of specific genomic aberrations.

Early genomic changes are presumably beneficial for tumor initiation[27], but it is unclear the effect of late truncal events. Here we have provided evidence that gains occurring later in the somatic evolution, i.e., close to MRCA, can also be beneficial. The simplified two-event cancer development model posits that the cancer-initiating event is followed by the promoting event [50].

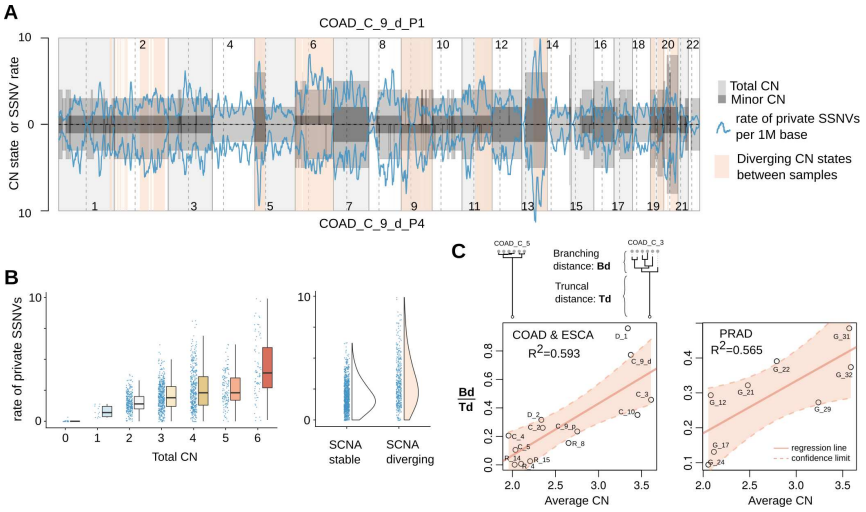

Fig. 6 The effect of SCNA on SSNV diversification during tumor expansion. (A) The rate of sample-private SSNVs when comparing two samples of a COAD patient tumor. The segmental CN states (total and minor CN) along the autosomal chromosomes for the two tumor samples are shown as gray rectangles above and below the x axis. The rate of sample private SSNVs (per million base pairs, blue line) fluctuates with the CN states. Genomic regions with different CN states between the two samples are in light red background. (B) The box plots on the left panel illustrate the rate of private SSNVs in sample P1 detected in regions at a given total CN state. The half-violin plots on the right panel demonstrate such a rate for regions showing stable or diverging CN states between the two samples. (C) The branching distance relative to truncal distance in a tumor's phylogenetic tree was calculated for each of the COAD, ESCA and PRAD tumors to evaluate the correlation with the averaged CN of the corresponding tumor samples. Annotations show the percentage of variance explained by a linear regression model.

We reason that the evolving CN gains might render the progenitor cell capable of "self-promoting," as they act similarly as a tumor "promoter" by (a) increasing the dosage of genes causing sustained proliferative signaling; (b) amplifying the mutant allele with early initiating driver variants; and (c) accelerating the accumulation of further genomic alterations. As both the early and late CN alterations could confer fitness advantages, chromosomal regions with SCNAs initiated early and arrived late, i.e., showing repetitive gains accompanying the entire course of the somatic evolution, could function as copy number "addictions."

GD, a landmark event in CN evolution, has context-dependent fitness effect. The punctuated CN gains successfully induced the SRE in tumors that underwent a late GD. By contrast, for many other tumors, especially osteosarcoma, GD was followed by additional CN gains that produces the MRCA. GD could tolerate the occurrence of deleterious passengers [51]. However, simply escaping purifying selection was not sufficient to drive the ultimate outgrowth, at least not in the tumors with post-GD gains, where some chromosomal regions can reach higher CN states. Alternatively, GD may create an inflated genome space, accelerating the accumulation of driver alterations. As GD itself

affects many genes, regions with pre- and/or post-GD gains could serve as a reduced search space for CN drivers.

Our method is applicable to a wide range of SCNAs, yet it is still challenging to analyze extremely high CN states (i.e. above eight). We note that regions with such a high CN likely evolve over time, such as the unequal segregation of extra-chromosomal oncogene amplifications [52, 53]. As such, late-arriving changes are expected for these amplified regions. Some focal high-level gains could involve small segments where the number of SSNVs is inadequate for calculation. This problem can be mitigated by borrowing information from nearby segments with the identical CN state. This strategy is applicable to synchronized SCNAs, such as chromothripsis [54, 55]. In addition, our analysis may have missed some late-appearing SCNAs due to overestimation. This is particularly evident for bi-allelic gains (4:2), where the upper bounds overestimate the actual timing (Supplementary Figure 1). Furthermore, deletion was not modelled as it is unidentifiable [18]; by comparing CN profiles between subpopulations, however, it is possible to study deletions during tumor expansion.

Our findings also illustrate the existence of a fundamental connection between CN evolution and SSNV diversity, which can explain the positive correlation between aneuploidy and mutational burden when excluding hypermutated tumors [56, 57]. Such a connection also indicates that we need to account for the dynamic nature of ongoing SCNAs when measuring subclonal evolution, which remains a challenge [58]. Finally, our results suggest that much can be gained by including the SCNA arrival time in studying tumor evolution, thereby shifting focus on exclusively early drivers to the evolving genomic events that affect the rate of tumor progression.

4 Methods

Table 1 WGS data included in this study

Tumor	Ref.	Accession Code	Sampling[1]	#Samples[2]	#Patients
OS	Negri[28]	EGAD00001004482	MTS	17	9
	Chen[29]	EGAS00001000263	Single,MTS	24	22
COAD	Cross[16]	EGAD00001004966	MRS	43	7
	Dang[26]	phs001722.v1.p1	MRS,MTS	7	2
BRCA	Yates[30]	EGAD00001002696	MTS	26	12
	Kawazu[31]	JGAD000095	Single	10	10
PRAD	Gundem[33]	EGAD00001000891	MRS,MTS	47	9
ESCA	Ross-Innes[59]	EGAD00001001394	MRS	11	4

[1]MRS: multi-region sampling; MTS: multi-tumor sampling

[2]Samples passed our quality assessment (Supplementary Figure 7, 8 and 9).

4.1 Somatic variant calling from WGS data

Raw WGS data in bam or fastq formats were downloaded from public databases provided by the original publications (Table 1). The cumulative read depth distribution along the human genome (hg38) and the tumor purity and ploidy for each sample are illustrated in Supplementary Figure 7, 8, and 9. We have extended our existing pipeline, which had achieved a balance in sensitivity and specificity in detecting SSNVs by borrowing information across multiple samples[60], to allow the detection of clonal SCNAs and the breakpoints of structural variants.

SSNVs and INDELs: Analysis-ready read alignment bam files (against hg38) were generated according to the best practices, including indel realignment, base recalibration and flagging of duplicated reads. Raw candidate variants were produced by MuTect (v1.1.7) [3]. To reduce the false-positive rate due to misalignments or other technical artifacts and to salvage the variants that may be missed due to uneven read coverage between samples, the alignment features surrounding each candidate variant were collected for each sample. The heuristic-based criterion for the read alignment patterns was adopted to refine and variant calls as detailed previously [60]. Small insertions and deletions were called by using Strelka (v1.0.15)[61].

SCNAs: Copy number and tumor purity were estimated by using TitanCNA (v1.26.0)[25]. Germline heterozygous SNVs used as input to TitanCNA were identified using Samtools (v1.5)[62] and subject to the same filtering strategy as was applied to SSNVs. The one-clone solution reported by TitanCNA (i.e., the sample is dominated by a clone with an SCNA profile along the genome) globally fit the data of the read coverage and allelic imbalance well, with a few exceptions for which the two-clones solution are necessary to explain the data of specific genomic regions.

SVs: We incorporated two distinct SV calling tools relying on orthogonal approaches, i.e., DELLY (v0.7.8, abnormal read pair and split-read analysis)[63] and GRIDSS (v2.10.1, local assembly based algorithm)[64]. We focused on the SV breakpoints found by both tools, as these shared calls generally have higher quality (e.g. with higher breakpoint confidence) than those unique to each tool (Supplementary Figure 10). SV breakpoints were annotated with AnnotSV[65].

4.2 Analysis of genomic divergence

SCNA divergence: When multi-samples are available for a patient, the truncal and private SCNAs were identified as follows: (1) we partitioned the genome into disjoint segments by considering all the SCNAs called from the samples of the patient; (2) for each segment, we calculated a generalized likelihood ratio statistics for the comparison between two samples. The statistics is the ratio of the values of the likelihood function (the probability of observing the read depth ratio and B-allele frequency for SNPs in the region) evaluated at the maximum likelihood estimation in the sub-model (two samples have the same CN profile) and at the maximum likelihood estimation in the full-model; and (3) the statistics converges weakly to a random variable with chi-square distribution and thus can be used to determine if a segment shows significantly different SCNA states between the two samples. The term "truncal SCNAs" refers to SCNAs that exhibit no difference in pairwise comparisons.

Sample phylogeny: We applied Treeomics (v1.7.13) [66] to construct sample phylogenies from SSNV data. Treeomics takes into account the uncertainty due to purity differences and variations of read depth on the SSNV loci to derive robust sample phylogenies.

Clonality, multiplicity of SSNVs and SV breakpoints: SSNVs were classified as public (present in all tumor cells) or private according to the criterion as previously described for multi-sampling data [60]. In individual samples, clonal SSNVs were identified as those with the 95% confidence interval of cancer cell fraction (CCF) covering 1 [60]. We focused on the public SSNVs (multi-sampling) and clonal SSNVs (single sampling) for the timing analysis. For SSNVs or SV breakpoints existing in an SCNA region, we applied a binomial model to calculate the maximum likelihood estimates of the number of segment copies containing that variant as previously described [21].

4.3 Inferring the arrival and initiation time of SCNAs

4.3.1 Allele state distribution of SSNVs for a clonal SCNA

For SSNV i in an SCNA region (with CN configuration of $N_t : N_b$ and $M \geq 10$ SSNVs in total), we obtained from WGS the read counts carrying the mutant allele m_i out of the total number of reads d_i. Expectation Maximization algorithm was used to estimate the proportion of SSNVs at each possible allele fraction, i.e. a vector q that gives the probability of randomly acquired SSNVs in this region having an purity-adjusted allele frequency ($f_i = a_j$) for each

possible allele state $\frac{j}{N_t}$. The rationale behind this is that the likelihood function of observing a particular SSNV data is related to the probability vector q as well as tumor purity. The log-likelihood is given by,

$$\sum_{i=1}^{M} \log \Pr(m_i \mid m_i > 0, q) = \sum_{i=1}^{M} \log \left(\frac{\sum_{j=1}^{N_t - N_b} \Pr(m_i \mid f_i = a_j) q_j}{1 - \sum_{j=1}^{N_t - N_b} (1 - a_j)^{d_i} q_j} \right). \quad (3)$$

4.3.2 Estimating the upper bounds of initiation and arrival time

For a genomic region affected by SCNA gains, let A be a possible history matrix with entry A_{jk} representing the number of segment copies in stage k that result in a final allele state a_j. Let q be a vector with entry q_j representing the probability of a randomly acquired SSNV in this region having allele frequency a_j. For single and double gains ($Nt : Nb$ at 3:1, 2:0, 4:1 or 3:0), the initiation time (t_0) and arrival time (t_K) are directly solved because matrix A is unique and invertible. For other SCNAs, `Butte` uses linear programming to obtain the upper bounds of timings across all possible history matrices for the corresponding CN configuration (Supplementary Figure 11). Let s denote the vector of the column sum of matrix A. Let t denote the relative time vector in which every component is a time fraction of the corresponding copy number state. We solve the following optimization problem by linear programming:

$$\begin{aligned} \max_{t_k} \quad & t_K \\ \text{s.t.} \quad & (A - qs^T)t = \mathbf{0} \\ & \mathbf{1} \cdot t = 1, \end{aligned}$$

where t_K is the last element in vector t. The maximum value of t_K gives us an upper bound of the arrival time given A. For upper bounds of initiation, we instead maximize t_0 which is the first element in t. To tolerate noise in the allele state distribution estimated from sequencing data, we add a slack variable on each capacity constraint, having a penalty cost of 100. The confidence intervals of the estimated upper bounds were calculated through bootstrapping the SSNV data.

4.4 Mathematical modeling of the late evolving gains

Consider two contrasting models based on multi-type branching processes with mutations. In both models, the tumor grows from a single tumor-initiating cell which just acquired GD. During the tumor's progression, cancer cells accumulate mutations (post-GD gains).

In the first model, all mutations are passenger mutations. Therefore, all cancer cells give birth at a rate of a_0 and die at a rate of b_0. The net growth

rate is $\lambda_0 = a_0 - b_0 > 0$. Neutral mutations occur at rate u_0 per unit time throughout the lifetime of a cell, and each mutation is distinct (the infinite-sites model of Kimura [67]). We can obtain that the number of mutations accumulated before the first cell division, which results in two descendants with infinite lineage, follows a geometric distribution with parameter $\frac{\lambda_0}{\lambda_0+u_0}$ and mean $\frac{u_0}{\lambda_0}$ (conditioned on the non-extinction of the cancer cell population). We then investigated the number of mutations which are shared by more than 90% of the total population (we refer to them as dominant mutations). Gunnarsson and his co-authors [41] derived exact expressions for the expected SFS of a cell population that evolves according to a branching process. We utilized their results on the skeleton subpopulation (see Appendix C of [41]) — cells with an infinite line of descents which determines the high frequency spectrum — to express the expected number of dominant mutations $\tilde{S}$ when the tumor reaches a fixed size N as

$$\tilde{S} = \frac{N}{\lceil 0.9N \rceil} \cdot \frac{u_0}{\lambda_0} \approx 1.11 \frac{u_0}{\lambda_0}. \tag{4}$$

In the second model, the tumor-initiating cell and its descendants with only passenger mutations form the type 0 population. Type 0 cells give birth at a rate of a_0 and die at a rate of b_0. The net growth rate is $\lambda_0 = a_0 - b_0 > 0$. Type 0 cells mutate to type 1 cells at a rate of u_1. Type 1 cells give birth at a rate of a_1 and die at a rate of b_1. The net growth rate is $\lambda_1 = a_1 - b_1 > \lambda_0$. Both type 0 and type 1 cells accumulate passenger mutations at a rate of u_0. Due to the selective advantage type 1 cells have over type 0 cells, a tumor at detection is most likely to be dominated by type 1 cells. As a result, it suffices to focus on mutations shared by the majority of type 1 cells. For tractability, we investigated the number of passenger mutations accrued in the first type 1 cell with infinite lineage. In Lemma 1, we obtained the distribution of the time to the first such type 1 cell conditioned on the non-extinction of the tumor.

Lemma 1 *Let σ_1 denote the time of occurrence of the first type 1 cell that gives rise to a family which does not die out, and let Ω_∞ denote the event of non-extinction of the tumor. Then*

$$\mathbb{P}\left(\sigma_1 > t \mid \Omega_\infty\right) = \frac{a_0\left(1-q_0\right) + \frac{u_1(1-q_1)}{1-q_0}}{a_0\left(1-q_0\right) + \frac{u_1(1-q_1)}{1-q_0} e^{\zeta t}},$$

where

$$q_0 = \frac{a_0 + b_0 + u_1 - \sqrt{\left(a_0 + b_0 + u_1\right)^2 - 4a_0\left(u_1 q_1 + b_0\right)}}{2a_0},$$

$$q_1 = \frac{b_1}{a_1}, \quad \textit{and}$$

$$\zeta = \frac{u_1\left(1-q_1\right)}{1-q_0} + a_0\left(1-q_0\right).$$

With Lemma 1, we can obtain the expected number of passenger mutations accumulated in the first type 1 cell with infinite lineage, denoted by $\bar{S}$:

$$\bar{S} = \int_0^\infty \mathbb{P}\left(\sigma_1 > t \mid \Omega_\infty\right) u_0 dt. \tag{5}$$

With (4) and (5), we obtained that the expected number of dominant post-GD gains in the subpopulation generated from the first type 1 cell with infinite lineage is $\tilde{S} + \bar{S} + 1$, where the last 1 represents the number of post-GD driver mutation. Proof for Lemma 1 and details of (4) can be found in Supplementary Methods.

Acknowledgments. This study uses the computing resources in Minnesota Supercomputing Institute. This manuscript was prepared using limited access datasets obtained from BC CANCER (OS_N), Queen Mary University of London (COAD_C), the Cancer Genome Project in Wellcome Sanger Institute (BRCA_Y), and does not necessarily reflect the opinions or views of the corresponding provider institutions. The OS_C dataset was generated by the St. Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project. The BRCA_K dataset was originally generated by research led by Dr. Masahito Kawazu and available at the website of the National Bioscience Database Center (NBDC; http://biosciencedbc.jp/en/) of the Japan Science and Technology Agency (JST). We thank International Cancer Genome Consortium (ICGC) for providing the access to the PRAD_G dataset (originally produced by Cancer Research UK Prostate Cancer Group Study) and ESCA_R dataset (by the Oesophageal Cancer Clinical and Molecular Stratification Study Group). We would also like to acknowledge the database of Genotypes and Phenotypes (dbGaP) and Dr. Ryan C. Fields for producing the COAD_D dataset (phs001722.v1.p1) which was supported by Siteman Cancer Center Investment Program. We thank Dr. Scott Dehm and Boyang Liu for commenting on the manuscript. We thank Einar Gunnarsson for sharing his proof ideas on the skeleton of the branching process.

Code availability. Code for `Butte` (computational framework to estimate SCNA arrival and initiating time from WGS data) and the mathematical modeling is available at https://github.com/SunPathLab/Butte/

Author contributions. R.S. and Z.W. designed the study. R.S., Y.X., and Z.W. developed the algorithms. Z.W., R.S., and A.N.N. constructed mathematical models and performed simulation studies. R.S., Y.X., Z.W. and L.M. performed the analysis of WGS data and visualized the results. R.S., Z.W., Y.X., N.M. and J.M.S interpreted the results and wrote the manuscript. All authors reviewed and provided feedback on the manuscript.

References

[1] Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: The next generation. Cell **144**(5), 646–674 (2011) arXiv:0208024 [gr-qc]. https://doi.org/10.1016/j.cell.2011.02.013

[2] Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., Van Loo, P., Ju, Y.S., Smid, M., Brinkman, A.B., Morganella, S., Aure, M.R., Lingjærde, O.C., Langerød, A., Ringnér, M., Ahn, S.-M., Boyault, S., Brock, J.E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J.A., Gerstung, M., Hooijer, G.K.J., Jang, S.J., Jones, D.R., Kim, H.-Y., King, T.A., Krishnamurthy, S., Lee, H.J., Lee, J.-Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C.A., Raine, K., Ramakrishnan, K., Rodríguez-González, F.G., Romieu, G., Sieuwerts, A.M., Simpson, P.T., Shepherd, R., Stebbings, L., Stefansson, O.A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G.G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B.K.T., Jonkers, J., Borg, Å., Ueno, N.T., Sotiriou, C., Viari, A., Futreal, P.A., Campbell, P.J., Span, P.N., Van Laere, S., Lakhani, S.R., Eyfjord, J.E., Thompson, A.M., Birney, E., Stunnenberg, H.G., van de Vijver, M.J., Martens, J.W.M., Børresen-Dale, A.-L., Richardson, A.L., Kong, G., Thomas, G., Stratton, M.R.: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature **534**(7605), 47–54 (2016). https://doi.org/10.1038/nature17676

[3] Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology **31**(3), 213–219 (2013) arXiv:NIHMS150003. https://doi.org/10.1038/nbt.2514

[4] Vasudevan, A., Schukken, K.M., Sausville, E.L., Girish, V., Adebambo, O.A., Sheltzer, J.M.: Aneuploidy as a promoter and suppressor of malignant growth. Nature Reviews Cancer 2021 21:2 **21**(2), 89–103 (2021). https://doi.org/10.1038/s41568-020-00321-1

[5] Hastings, P.J., Lupski, J.R., Rosenberg, S.M., Ira, G.: Mechanisms of change in gene copy number. Nature Reviews Genetics **10**(8), 551–564 (2009). https://doi.org/10.1038/nrg2593

[6] Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., Berney, D., Scher, H., Carver, B., Scardino, P., Schultz, N., Taylor, B., Vickers, A., Cuzick, J., Sawyers, C.L.: Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and

death. eLife **7** (2018). https://doi.org/10.7554/ELIFE.37294

[7] Nowak, M.A., Komarova, N.L., Sengupta, A., Jallepalli, P.V., Shih, L.M., Vogelstein, B., Lengauer, C.: The role of chromosomal instability in tumor initiation. Proceedings of the National Academy of Sciences of the United States of America **99**(25), 16226–16231 (2002). https://doi.org/10.1073/pnas.202617399

[8] Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., Elledge, S.J.: XCumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell **155**(4), 948–62 (2013). https://doi.org/10.1016/j.cell.2013.10.011

[9] Lukow, D.A., Sausville, E.L., Suri, P., Chunduri, N.K., Leu, J., Kendall, J., Wang, Z., Storchova, Z., Sheltzer, J.M.: Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. bioRxiv, 2020–0925314229 (2020). https://doi.org/10.1101/2020.09.25.314229

[10] Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., Wala, J., Mermel, C.H., Sougnez, C., Gabriel, S.B., Hernandez, B., Shen, H., Laird, P.W., Getz, G., Meyerson, M., Beroukhim, R.: Pan-cancer patterns of somatic copy number alteration. Nature Genetics **45**(10), 1134–1140 (2013) arXiv:15334406. https://doi.org/10.1038/ng.2760

[11] Roberts, R.D., Lizardo, M.M., Reed, D.R., Hingorani, P., Glover, J., Allen-Rhoades, W., Fan, T., Khanna, C., Sweet-Cordero, E.A., Cash, T., Bishop, M.W., Hegde, M., Sertil, A.R., Koelsche, C., Mirabello, L., Malkin, D., Sorensen, P.H., Meltzer, P.S., Janeway, K.A., Gorlick, R., Crompton, B.D.: Provocative questions in osteosarcoma basic and translational biology: A report from the Children's Oncology Group. John Wiley and Sons Inc. (2019). https://doi.org/10.1002/cncr.32351. http://www.ncbi.nlm.nih.gov/pubmed/31355930http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6948723

[12] Jolly, C., Van Loo, P.: Timing somatic events in the evolution of cancer. Genome Biology **19**(1), 95 (2018). https://doi.org/10.1186/s13059-018-1476-3

[13] Nowell, P.: The clonal evolution of tumor cell populations. Science **194**(4260), 23–28 (1976). https://doi.org/10.1126/science.959840

[14] Caldas, C.: Cancer sequencing unravels clonal evolution. Nature Biotechnology **30**(5), 408–410 (2012). https://doi.org/10.1038/nbt.2213

[15] Watkins, T.B.K., Lim, E.L., Petkovic, M., Elizalde, S., Birkbak, N.J., Wilson, G.A., Moore, D.A., Grönroos, E., Rowan, A., Dewhurst, S.M.,

Demeulemeester, J., Dentro, S.C., Horswell, S., Au, L., Haase, K., Escudero, M., Rosenthal, R., Bakir, M.A., Xu, H., Litchfield, K., Lu, W.T., Mourikis, T.P., Dietzen, M., Spain, L., Cresswell, G.D., Biswas, D., Lamy, P., Nordentoft, I., Harbst, K., Castro-Giner, F., Yates, L.R., Caramia, F., Jaulin, F., Vicier, C., Tomlinson, I.P.M., Brastianos, P.K., Cho, R.J., Bastian, B.C., Dyrskjøt, L., Jönsson, G.B., Savas, P., Loi, S., Campbell, P.J., Andre, F., Luscombe, N.M., Steeghs, N., Tjan-Heijnen, V.C.G., Szallasi, Z., Turajlic, S., Jamal-Hanjani, M., Van Loo, P., Bakhoum, S.F., Schwarz, R.F., McGranahan, N., Swanton, C.: Pervasive chromosomal instability and karyotype order in tumour evolution. Nature, 1–7 (2020). https://doi.org/10.1038/s41586-020-2698-6

[16] Cross, W., Kovac, M., Mustonen, V., Temko, D., Davis, H., Baker, A.-M., Biswas, S., Arnold, R., Chegwidden, L., Gatenbee, C., Anderson, A.R., Koelzer, V.H., Martinez, P., Jiang, X., Domingo, E., Woodcock, D.J., Feng, Y., Kovacova, M., Maughan, T., Jansen, M., Rodriguez-Justo, M., Ashraf, S., Guy, R., Cunningham, C., East, J.E., Wedge, D.C., Wang, L.M., Palles, C., Heinimann, K., Sottoriva, A., Leedham, S.J., Graham, T.A., Tomlinson, I.P.M.: The evolutionary landscape of colorectal tumorigenesis. Nature Ecology Evolution **2**, 1661–1672 (2018). https://doi.org/10.1038/s41559-018-0642-z

[17] Reiter, J.G., Makohon-Moore, A.P., Gerold, J.M., Heyde, A., Attiyeh, M.A., Kohutek, Z.A., Tokheim, C.J., Brown, A., DeBlasio, R.M., Niyazov, J., Zucker, A., Karchin, R., Kinzler, K.W., Iacobuzio-Donahue, C.A., Vogelstein, B., Nowak, M.A.: Minimal functional driver gene heterogeneity among untreated metastases. Science (New York, N.Y.) **361**(6406), 1033–1037 (2018). https://doi.org/10.1126/science.aat7171

[18] Purdom, E., Ho, C., Grasso, C.S., Quist, M.J., Cho, R.J., Spellman, P., Kelso, J.: Methods and challenges in timing chromosomal abnormalities within cancer samples. Bioinformatics **29**(24), 3113–3120 (2013). https://doi.org/10.1093/bioinformatics/btt546

[19] Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D.G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P.C., Bowtell, D.D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S.C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q.D., Dentro, S.C., Leshchiner, I., Gerstung, M., Jolly, C., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Gonzalez, S., Rubanova, Y., Macintyre, G., Adams, D.J., Anur, P., Beroukhim, R., Boutros, P.C., Bowtell, D.D., Campbell, P.J., Cao, S., Christie, E.L., Cmero, M., Cun,

Y., Dawson, K.J., Donmez, N., Drews, R.M., Eils, R., Fan, Y., Fittall, M., Garsed, D.W., Getz, G., Ha, G., Imielinski, M., Jerman, L., Ji, Y., Kleinheinz, K., Lee, J., Lee-Six, H., Livitz, D.G., Malikic, S., Markowetz, F., Martincorena, I., Mitchell, T.J., Mustonen, V., Oesper, L., Peifer, M., Peto, M., Raphael, B.J., Rosebrock, D., Sahinalp, S.C., Salcedo, A., Schlesner, M., Schumacher, S., Sengupta, S., Shi, R., Shin, S.J., Spiro, O., Stein, L.D., Vázquez-García, I., Vembu, S., Wheeler, D.A., Yang, T.P., Yao, X., Yuan, K., Zhu, H., Wang, W., Morris, Q.D., Spellman, P.T., Wedge, D.C., Van Loo, P., Spellman, P.T., Wedge, D.C.: The evolutionary history of 2,658 cancers. Nature **578**(7793), 122–128 (2020). https://doi.org/10.1038/s41586-019-1907-7

[20] Sheltzer, J.M., Ko, J.H., Replogle, J.M., Burgos, N.C.H., Chung, E.S., Meehl, C.M., Sayles, N.M., Passerini, V., Storchova, Z., Amon, A.: Single-chromosome Gains Commonly Function as Tumor Suppressors. Cancer Cell **31**(2), 240–255 (2017). https://doi.org/10.1016/J.CCELL.2016.12.004

[21] Greenman, C.D., Pleasance, E.D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K.W., Carter, N., Edwards, P.A.W., Futreal, P.A., Stratton, M.R., Campbell, P.J.: Estimation of rearrangement phylogeny for cancer genomes. Genome Research **22**(2), 346–361 (2012). https://doi.org/10.1101/gr.118414.110

[22] Dewhurst, S.M., McGranahan, N., Burrell, R.A., Rowan, A.J., Gr??nroos, E., Endesfelder, D., Joshi, T., Mouradov, D., Gibbs, P., Ward, R.L., Hawkins, N.J., Szallasi, Z., Sieber, O.M., Swanton, C.: Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discovery **4**(2), 175–185 (2014). https://doi.org/10.1158/2159-8290.CD-13-0285

[23] Storchova, Z., Kuffer, C.: The consequences of tetraploidy and aneuploidy. Journal of cell science **121**(Pt 23), 3859–66 (2008). https://doi.org/10.1242/jcs.039537

[24] Laughney, A.M., Elizalde, S., Genovese, G., Bakhoum, S.F.: Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution. Cell Reports **12**(5), 809–820 (2015). https://doi.org/10.1016/j.celrep.2015.06.065

[25] Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L.M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M.A., Gilks, C.B., Huntsman, D.G., McAlpine, J.N., Aparicio, S., Shah, S.P.: TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Research **24**(11), 1881–1893 (2014).

https://doi.org/10.1101/gr.180281.114

[26] Dang, H.X., Krasnick, B.A., White, B.S., Grossman, J.G., Strand, M.S., Zhang, J., Cabanski, C.R., Miller, C.A., Fulton, R.S., Goedegebuure, S.P., Fronick, C.C., Griffith, M., Larson, D.E., Goetz, B.D., Walker, J.R., Hawkins, W.G., Strasberg, S.M., Linehan, D.C., Lim, K.H., Lockhart, A.C., Mardis, E.R., Wilson, R.K., Ley, T.J., Maher, C.A., Fields, R.C.: The clonal evolution of metastatic colorectal cancer. Science Advances **6**(24), 9691 (2020). https://doi.org/10.1126/sciadv.aay9691

[27] Paterson, C., Clevers, H., Bozic, I.: Mathematical model of colorectal cancer initiation. Proceedings of the National Academy of Sciences of the United States of America **117**(34), 20681–20688 (2020). https://doi.org/10.1073/pnas.2003771117

[28] Negri, G.L., Grande, B.M., Delaidelli, A., El-Naggar, A., Cochrane, D., Lau, C.C., Triche, T.J., Moore, R.A., Jones, S.J., Montpetit, A., Marra, M.A., Malkin, D., Morin, R.D., Sorensen, P.H.: Integrative genomic analysis of matched primary and metastatic pediatric osteosarcoma. The Journal of Pathology **249**(3), 319–331 (2019). https://doi.org/10.1002/path.5319

[29] Chen, X., Bahrami, A., Pappo, A., Easton, J., Dalton, J., Hedlund, E., Ellison, D., Shurtleff, S., Wu, G., Wei, L., Parker, M., Rusch, M., Nagahawatte, P., Wu, J., Mao, S., Boggs, K., Mulder, H., Yergeau, D., Lu, C., Ding, L., Edmonson, M., Qu, C., Wang, J., Li, Y., Navid, F., Daw, N.C., Mardis, E.R., Wilson, R.K., Downing, J.R., Zhang, J., Dyer, M.A., St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project: Recurrent Somatic Structural Variations Contribute to Tumorigenesis in Pediatric Osteosarcoma. Cell Reports **7**(1), 104–112 (2014). https://doi.org/10.1016/j.celrep.2014.03.003

[30] Yates, L.R., Knappskog, S., Wedge, D., Farmery, J.H.R., Gonzalez, S., Martincorena, I., Alexandrov, L.B., Van Loo, P., Haugland, H.K., Lilleng, P.K., Gundem, G., Gerstung, M., Pappaemmanuil, E., Gazinska, P., Bhosle, S.G., Jones, D., Raine, K., Mudie, L., Latimer, C., Sawyer, E., Desmedt, C., Sotiriou, C., Stratton, M.R., Sieuwerts, A.M., Lynch, A.G., Martens, J.W., Richardson, A.L., Tutt, A., Lønning, P.E., Campbell, P.J.: Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell **32**(2), 169–1847 (2017). https://doi.org/10.1016/j.ccell.2017.07.005

[31] Kawazu, M., Kojima, S., Ueno, T., Totoki, Y., Nakamura, H., Kunita, A., Qu, W., Yoshimura, J., Soda, M., Yasuda, T., Hama, N., Saito-Adachi, M., Sato, K., Kohsaka, S., Sai, E., Ikemura, M., Yamamoto, S., Ogawa, T., Fukayama, M., Tada, K., Seto, Y., Morishita, S., Hazama, S., Shibata, T., Yamashita, Y., Mano, H.: Integrative analysis of genomic

alterations in triple-negative breast cancer in association with homologous recombination deficiency. PLOS Genetics **13**(6), 1006853 (2017). https://doi.org/10.1371/journal.pgen.1006853

[32] Ross-Innes, C.S., Becq, J., Warren, A., Cheetham, R.K., Northen, H., O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., Weaver, J.M.J., Lynch, A.G., Kingsbury, Z., Ross, M., Humphray, S., Bentley, D., Fitzgerald, R.C., Hayes, S.J., Ang, Y., Welch, I., Preston, S., Oakes, S., Save, V., Skipworth, R., Tucker, O., Davies, J., Crichton, C., Schusterreiter, C., Underwood, T., Noble, F., Stacey, B., Kelly, J., Byrne, J., Haydon, A., Sharland, D., Owsley, J., Barr, H., Lagergren, J., Gossage, J., Davies, A., Mason, R., Chang, F., Zylstra, J., Sanders, G., Wheatley, T., Berrisford, R., Bracey, T., Harden, C., Bunting, D., Roques, T., Nobes, J., Loo, S., Lewis, M., Cheong, E., Priest, O., Parsons, S.L., Soomro, I., Kaye, P., Saunders, J., Pang, V., Welch, N.T., Catton, J.A., Duffy, J.P., Ragunath, K., Lovat, L., Haidry, R., Miah, H., Kerr, S., Eneh, V., Butawan, R., Roques, T., Lewis, M., Cheong, E., Kumar, B., Igali, L., Walton, S., Dann, A., Safranek, P., Hindmarsh, A., Sudjendran, V., Scott, M., Cluroe, A., Miremadi, A., Mahler-Araujo, B., Nutzinger, B., Peters, C., Abdullahi, Z., Crawte, J., MacRae, S., Noorani, A., Elliott, R.F., Bower, L., Edwards, P., Tavare, S., Eldridge, M., Bornschein, J., Secrier, M., Yang, T.-P., O'Neill, J.R., Adamczuk, K., Lao-Sirieix, P., Grehan, N., Smith, L., Lishman, S., Beardsmore, D., Dawson, S.: Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. Nature Genetics **47**(9), 1038–1046 (2015). https://doi.org/10.1038/ng.3357

[33] Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M.C., Papaemmanuil, E., Brewer, D.S., Kallio, H.M.L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K.J., Isaacs, W., Emmert-Buck, M.R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H.C., Neal, D.E., Cooper, C.S., Eeles, R.A., Visakorpi, T., Campbell, P.J., McDermott, U., Wedge, D.C., Bova, G.S.: The evolutionary history of lethal metastatic prostate cancer. Nature **520**(7547), 353–357 (2015). https://doi.org/10.1038/nature14347

[34] Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G.: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology **12**(4), 41 (2011). https://doi.org/10.1186/GB-2011-12-4-R41

[35] Dewhurst, S.M., McGranahan, N., Burrell, R.A., Rowan, A.J., Grönroos, E., Endesfelder, D., Joshi, T., Mouradov, D., Gibbs, P., Ward, R.L., Hawkins, N.J., Szallasi, Z., Sieber, O.M., Swanton, C.: Tolerance of

whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer discovery **4**(2), 175–185 (2014). https://doi.org/10.1158/2159-8290.CD-13-0285

[36] Bielski, C.M., Zehir, A., Penson, A.V., Donoghue, M.T.A., Chatila, W., Armenia, J., Chang, M.T., Schram, A.M., Jonsson, P., Bandlamudi, C., Razavi, P., Iyer, G., Robson, M.E., Stadler, Z.K., Schultz, N., Baselga, J., Solit, D.B., Hyman, D.M., Berger, M.F., Taylor, B.S.: Genome doubling shapes the evolution and prognosis of advanced cancers. Nature Genetics **50**(8), 1189–1195 (2018). https://doi.org/10.1038/s41588-018-0165-1

[37] Sayles, L.C., Breese, M.R., Koehne, A.L., Leung, S.G., Lee, A.G., Liu, H.-Y., Spillinger, A., Shah, A.T., Tanasa, B., Straessler, K., Hazard, F.K., Spunt, S.L., Marina, N., Kim, G.E., Cho, S.-J., Avedian, R.S., Mohler, D.G., Kim, M.-O., Dubois, S.G., Hawkins, D.S., Sweet-Cordero, E.A.: Genome-Informed Targeted Therapy for Osteosarcoma. Cancer Discovery **9**(1), 17–1152 (2018). https://doi.org/10.1158/2159-8290.CD-17-1152

[38] N, L., D, L., X, H., W, L., G, L., Q, W., T, Z., J, G., Y, L., Q, Z.: RUNX2 and Osteosarcoma. Anti-cancer agents in medicinal chemistry **15**(7), 881–887 (2015). https://doi.org/10.2174/1871520615666150304151228

[39] Zhang, C., Liu, J., Xu, D., Zhang, T., Hu, W., Feng, Z.: Gain-of-function mutant p53 in cancer progression and therapy. Journal of Molecular Cell Biology **12**(9), 674–687 (2020). https://doi.org/10.1093/JMCB/MJAA040

[40] Durrett, R.: Branching Process Models of Cancer. In: Branching Process Models of Cancer. Mathematical Biosciences Institute Lecture Series, Vol 1.1, pp. 1–63. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-16065-8_1

[41] Gunnarsson, E.B., Leder, K., Foo, J.: Exact site frequency spectra of neutrally evolving tumors: a transition between power laws reveals a signature of cell viability. Theoretical Population Biology (2021) arXiv:2102.11959

[42] Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Westra, H.J., Maloney, D., Simeonov, A., Pers, T.H., Hirschhorn, J.N., Jansen, R.C., Schultes, E.A., Van Haagen, H.H.H.B.M., De Vries, E.G.E., Te Meerman, G.J., Wijmenga, C., Van Vugt, M.A.T.M., Franke, L.: Gene expression analysis identifies global gene dosage sensitivity in cancer. Nature Genetics 2015 47:2 **47**(2), 115–125 (2015). https://doi.org/10.1038/ng.3173

[43] Rice, A.M., McLysaght, A.: Dosage-sensitive genes in evolution and disease. BMC Biology 2017 15:1 **15**(1), 1–10 (2017). https://doi.org/10.1186/S12915-017-0418-Y

[44] Rice, A.M., McLysaght, A.: Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nature Communications 2017 8:1 **8**(1), 1–11 (2017). https://doi.org/10.1038/ncomms14366

[45] Edwards, Z.C., Trotter, E.W., Torres-Ayuso, P., Chapman, P., Wood, H.M., Nyswaner, K., Brognard, J.: Survival of Head and Neck Cancer Cells Relies upon LZK Kinase-Mediated Stabilization of Mutant p53. Cancer Research **77**(18), 4961–4972 (2017). https://doi.org/10.1158/0008-5472.CAN-17-0267

[46] Han, H., Chen, Y., Cheng, L., Prochownik, E.V., Li, Y., Han, H., Chen, Y., Cheng, L., Prochownik, E.V., Li, Y.: microRNA-206 impairs c-Myc-driven cancer in a synthetic lethal manner by directly inhibiting MAP3K13. Oncotarget **7**(13), 16409–16419 (2016). https://doi.org/10.18632/ONCOTARGET.7653

[47] Zhang, L., Shay, J.W.: Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. JNCI Journal of the National Cancer Institute **109**(8) (2017). https://doi.org/10.1093/JNCI/DJW332

[48] Waters, C.E., Saldivar, J.C., Hosseini, S.A., Huebner, K.: The FHIT gene product: tumor suppressor and genome 'caretaker'. Cellular and molecular life sciences : CMLS **71**(23), 4577 (2014). https://doi.org/10.1007/S00018-014-1722-0

[49] Sakthianandeswaren, A., Parsons, M.J., Mouradov, D., Mackinnon, R.N., Catimel, B., Liu, S., Palmieri, M., Love, C., Jorissen, R.N., Li, S., Whitehead, L., Putoczki, T.L., Preaudet, A., Tsui, C., Nowell, C.J., Ward, R.L., Hawkins, N.J., Desai, J., Gibbs, P., Ernst, M., Street, I., Buchert, M., Sieber, O.M.: MACROD2 Haploinsufficiency Impairs Catalytic Activity of PARP1 and Promotes Chromosome Instability and Growth of Intestinal Tumors. Cancer Discovery **8**(8), 988–1005 (2018). https://doi.org/10.1158/2159-8290.CD-17-0909

[50] Moolgavkar, S.H., Venzon, D.J.: Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. Mathematical Biosciences **47**(1-2), 55–77 (1979). https://doi.org/10.1016/0025-5564(79)90005-1

[51] López, S., Lim, E.L., Horswell, S., Haase, K., Huebner, A., Dietzen, M., Mourikis, T.P., Watkins, T.B.K., Rowan, A., Dewhurst, S.M., Birkbak, N.J., Wilson, G.A., Van Loo, P., Jamal-Hanjani, M., Swanton, C., McGranahan, N.: Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nature Genetics **52**(3), 283–293 (2020). https://doi.org/10.1038/s41588-020-0584-7

[52] Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee,

C., Li, B., Arden, K., Ren, B., Nathanson, D.A., Kornblum, H.I., Taylor, M.D., Kaushal, S., Cavenee, W.K., Wechsler-Reya, R., Furnari, F.B., Vandenberg, S.R., Rao, P.N., Wahl, G.M., Bafna, V., Mischel, P.S.: Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature **543**(7643), 122–125 (2017) arXiv:NIHMS150003. https://doi.org/10.1038/nature21356

[53] Kim, H., Nguyen, N.P., Turner, K., Wu, S., Gujar, A.D., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., Amin, S.B., Yi, E., Menghi, F., Schulte, J.H., Henssen, A.G., Chang, H.Y., Beck, C.R., Mischel, P.S., Bafna, V., Verhaak, R.G.W.: Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nature Genetics 2020 52:9 **52**(9), 891–897 (2020). https://doi.org/10.1038/s41588-020-0678-2

[54] Cortés-Ciriano, I., Lee, J.J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., Park, P.J.: Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nature Genetics 2020 52:3 **52**(3), 331–341 (2020). https://doi.org/10.1038/s41588-019-0576-7

[55] Zhang, C.-Z., Leibowitz, M.L., Pellman, D.: Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes development **27**(23), 2513–30 (2013). https://doi.org/10.1101/gad.229559.113

[56] Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., Lazar, A.J., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., Kasapi, M., Ferguson, M.L., Hutter, C.M., Sofia, H.J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J.C., Zhang, J.J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D.I., Kim, J., Lawrence, M.S., Lin, P., Meier, S., Noble, M.S., Saksena, G., Voet, D., Zhang, H., Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B.M., Hegde, A.M., Ju, Z., Kanchi, R.S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G.B., Ng, K.S., Rao, A., Ryan, M., Wang, J., Weinstein, J.N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W.K., de Bruijn, I., Gao, J., Gross, B.E., Heins, Z.J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M.G., Ochoa, A., Phillips, S.M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S.O., Sun, Y., Taylor, B.S., Wang, J., Zhang, H., Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J.M., Wong, C.K., Yau, C., Hayes, D.N., Parker, J.S., Wilkerson, M.D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S.J.M.,

Kasaian, K., Lee, D., Ma, Y., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Mungall, K., Robertson, A.G., Sadeghi, S., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A.C., Beroukhim, R., Cherniack, A.D., Cibulskis, C., Gabriel, S.B., Ha, G., Meyerson, M., Schumacher, S.E., Shih, J., Kucherlapati, M.H., Kucherlapati, R.S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M.S., Lai, P.H., Maglinte, D.T., Van Den Berg, D.J., Weisenberger, D.J., Auman, J.T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K.A., Hoyle, A.P., Jefferys, S.R., Jones, C.D., Meng, S., Mieczkowski, P.A., Mose, L.E., Perou, A.H., Perou, C.M., Roach, J., Shi, Y., Simons, J.V., Skelly, T., Soloway, M.G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P.W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C.J., Dinh, H., Doddapaneni, H.V., Donehower, L.A., Drummond, J., Gibbs, R.A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbrot, E., Wang, L., Wang, M., Wheeler, D.A., Xi, L., Zhao, F., Hess, J., Appelbaum, E.L., Bailey, M., Cordes, M.G., Ding, L., Fronick, C.C., Fulton, L.A., Fulton, R.S., Kandoth, C., Mardis, E.R., McLellan, M.D., Miller, C.A., Schmidt, H.K., Wilson, R.K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J.M., Gerken, M., Leraas, K.M., Lichtenberg, T.M., Ramirez, N.C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A.L., de Carvalho, A.C., Fregnani, J.H., Longatto-Filho, A., Reis, R.M., Scapulatempo-Neto, C., Silveira, H.C.S., Vidal, D.O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M.L., Castro, P.D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E.R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C.P., Malykh, A., Barnholtz-Sloan, J.S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q.T., Shimmel, K., Wolinsky, Y., Sloan, A.E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B.Y., Hagedorn, C.H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L.A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R.J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S.M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M.H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D.J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J.J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D.M., Sica, G., Van Meir, E.G., Zhang, H., Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., van Kessel, K.E., Zwarthoff, E.C.,

Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W.J., Martin, J., Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W.G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A.L., Van Bang, N., Hanh, P.T., Phu, B.D., Tang, Y., Colman, H., Evason, K., Dottino, P.R., Martignetti, J.A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K.J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A.H., Castle, E., Chandan, V., Cheville, J., Copland, J.A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J.P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B.P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R.H., Torbenson, M., Yang, J.D., Zhang, L., Brimo, F., Ajani, J.A., Angulo Gonzalez, A.M., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T.A., Ghossein, R., Gopalan, A., Levine, D.A., Reuter, V., Singer, S., Singh, B., Tien, N.V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J.W., Hung, N.P., Kebebew, E., Linehan, W.M., Metwalli, A.R., Pacak, K., Pinto, P.A., Schiffman, M., Schmidt, L.S., Vocke, C.D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D.M., Rintoul, R.C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., de Krijger, R., Gimenez-Roqueplo, A.P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N.A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J.A., Liptay, M.J., Pool, M., Seder, C.W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M.C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K.F., Janssen, K.P., Slotta-Huspenina, J., Abdel-Rahman, M.H., Aziz, D., Bell, S., Cebulla, C.M., Davis, A., Duell, R., Elder, J.B., Hilty, J., Kumar, B., Lang, J., Lehman, N.L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal,

K., Wakely, P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W.E., Sexton, K.C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S.L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P.R., Chan, J.M., Disaia, P., Glenn, P., Kelley, R.K., Landen, C.N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., von Deimling, A., Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A.K., Madan, R., Rosenthal, H.G., Adebamowo, C., Adebamowo, S.N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A.M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Mora Pinero, E.M., Quintero-Aguilo, M., Carlotti, C.G., Dos Santos, J.S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C.S., Godwin, E.M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K.M., Yang, I., Korst, R., Rathmell, W.K., Fantacone-Campbell, J.L., Hooke, J.A., Kovatich, A.J., Shriver, C.D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M.A., Lee, J.I., Aredes, N.D., Mariamidze, A., Cherniack, A.D., Beroukhim, R., Meyerson, M.: Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell **33**(4), 676–6893 (2018). https://doi.org/10.1016/j.ccell.2018.03.007

[57] Davoli, T., Uno, H., Wooten, E.C., Elledge, S.J.: Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science **355**(6322), 8399 (2017). https://doi.org/10.1126/science.aaf8399

[58] Satas, G., Zaccaria, S., Mon, G., Raphael, B.J.: SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses. Cell Systems **10**(4), 323–3328 (2020). https://doi.org/10.1016/J.CELS.2020.04.001

[59] Ross-Innes, C.S., Becq, J., Warren, A., Cheetham, R.K., Northen, H., O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., Weaver, J.M.J., Lynch, A.G., Kingsbury, Z., Ross, M., Humphray, S., Bentley, D., Fitzgerald, R.C., Hayes, S.J., Ang, Y., Welch, I., Preston, S., Oakes, S., Save, V., Skipworth, R., Tucker, O., Davies, J., Crichton, C., Schusterreiter, C., Underwood, T., Noble, F., Stacey, B., Kelly, J., Byrne, J., Haydon, A., Sharland, D., Owsley, J., Barr, H., Lagergren, J., Gossage, J., Davies, A., Mason, R., Chang, F., Zylstra, J., Sanders, G., Wheatley, T., Berrisford, R., Bracey, T., Harden, C., Bunting, D., Roques, T., Nobes, J., Loo,

S., Lewis, M., Cheong, E., Priest, O., Parsons, S.L., Soomro, I., Kaye, P., Saunders, J., Pang, V., Welch, N.T., Catton, J.A., Duffy, J.P., Ragunath, K., Lovat, L., Haidry, R., Miah, H., Kerr, S., Eneh, V., Butawan, R., Roques, T., Lewis, M., Cheong, E., Kumar, B., Igali, L., Walton, S., Dann, A., Safranek, P., Hindmarsh, A., Sudjendran, V., Scott, M., Cluroe, A., Miremadi, A., Mahler-Araujo, B., Nutzinger, B., Peters, C., Abdullahi, Z., Crawte, J., MacRae, S., Noorani, A., Elliott, R.F., Bower, L., Edwards, P., Tavare, S., Eldridge, M., Bornschein, J., Secrier, M., Yang, T.-P., O'Neill, J.R., Adamczuk, K., Lao-Sirieix, P., Grehan, N., Smith, L., Lishman, S., Beardsmore, D., Dawson, S.: Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. Nature Genetics **47**(9), 1038–1046 (2015). https://doi.org/10.1038/ng.3357

[60] Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., Fischer, J.M., Shibata, D., Curtis, C.: Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature Genetics **49**(7), 1015–1024 (2017). https://doi.org/10.1038/ng.3891

[61] Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., Cheetham, R.K.: Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics **28**(14), 1811–1817 (2012). https://doi.org/10.1093/bioinformatics/bts271

[62] Li, H.: Improving SNP discovery by base alignment quality. Bioinformatics **27**(8), 1157–1158 (2011). https://doi.org/10.1093/bioinformatics/btr076

[63] Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., Korbel, J.O.: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England) **28**(18), 333–339 (2012). https://doi.org/10.1093/bioinformatics/bts378

[64] Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., Papenfuss, A.T.: GRIDSS: Sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Research **27**(12), 2050–2060 (2017). https://doi.org/10.1101/gr.222109.117

[65] Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., Muller, J.: AnnotSV: an integrated tool for structural variations annotation. Bioinformatics **34**(20), 3572–3574 (2018). https://doi.org/10.1093/BIOINFORMATICS/BTY304

[66] Reiter, J.G., Makohon-Moore, A.P., Gerold, J.M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C.A., Vogelstein, B., Nowak, M.A.: Reconstructing metastatic seeding patterns of human cancers. Nature Communications **8**, 14114 (2017). https://doi.org/10.1038/ncomms14114

[67] Kimura, M.: The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. Genetics **61**(4), 893 (1969). https://doi.org/10.1093/genetics/61.4.893