Article

The evolutionary history of 2,658 cancers

https://doi.org/10.1038/s41586-019-1907-7

Received: 11 August 2017

Accepted: 18 November 2019

Published online: 5 February 2020

Open access

Moritz Gerstung[1,2,3,920*], Clemency Jolly[4,920], Ignaty Leshchiner[5,920], Stefan C. Dentro[3,4,6,920], Santiago Gonzalez[1,920], Daniel Rosebrock[5], Thomas J. Mitchell[3,7], Yulia Rubanova[8,9], Pavana Anur[10], Kaixian Yu[11], Maxime Tarabichi[3,4], Amit Deshwar[8,9], Jeff Wintersinger[8,9], Kortine Kleinheinz[12,13], Ignacio Vázquez-García[3,7], Kerstin Haase[4], Lara Jerman[1,14], Subhajit Sengupta[15], Geoff Macintyre[16], Salem Malikic[17,18], Nilgun Donmez[17,18], Dimitri G. Livitz[5], Marek Cmero[19,20], Jonas Demeulemeester[4,21], Steven Schumacher[5], Yu Fan[11], Xiaotong Yao[22,23], Juhee Lee[24], Matthias Schlesner[12], Paul C. Boutros[8,25,26], David D. Bowtell[27], Hongtu Zhu[11], Gad Getz[5,28,29,30], Marcin Imielinski[22,23], Rameen Beroukhim[5,31], S. Cenk Sahinalp[18,32], Yuan Ji[15,33], Martin Peifer[34], Florian Markowetz[16], Ville Mustonen[35], Ke Yuan[16,36], Wenyi Wang[11], Quaid D. Morris[8,9], PCAWG Evolution & Heterogeneity Working Group[37], Paul T. Spellman[10,921], David C. Wedge[6,38,921], Peter Van Loo[4,21,921*] & PCAWG Consortium[39]

Cancer develops through a process of somatic evolution[1,2]. Sequencing data from a single biopsy represent a snapshot of this process that can reveal the timing of specific genomic aberrations and the changing influence of mutational processes[3]. Here, by whole-genome sequencing analysis of 2,658 cancers as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)[4], we reconstruct the life history and evolution of mutational processes and driver mutation sequences of 38 types of cancer. Early oncogenesis is characterized by mutations in a constrained set of driver genes, and specific copy number gains, such as trisomy 7 in glioblastoma and isochromosome 17q in medulloblastoma. The mutational spectrum changes significantly throughout tumour evolution in 40% of samples. A nearly fourfold diversification of driver genes and increased genomic instability are features of later stages. Copy number alterations often occur in mitotic crises, and lead to simultaneous gains of chromosomal segments. Timing analyses suggest that driver mutations often precede diagnosis by many years, if not decades. Together, these results determine the evolutionary trajectories of cancer, and highlight opportunities for early cancer detection.

Similar to the evolution in species, the approximately 10^{14} cells in the human body are subject to the forces of mutation and selection[1]. This process of somatic evolution begins in the zygote and only comes to rest at death, as cells are constantly exposed to mutagenic stresses, introducing 1–10 mutations per cell division[2]. These mutagenic forces lead to a gradual accumulation of point mutations throughout life, observed in a range of healthy tissues[5–11] and cancers[12]. Although these mutations are predominantly selectively neutral passenger mutations, some are proliferatively advantageous driver mutations[13]. The types of mutation in cancer genomes are well studied, but little is known about the times when these lesions arise during somatic evolution and where the boundary between normal evolution and cancer progression should be drawn.

Sequencing of bulk tumour samples enables partial reconstruction of the evolutionary history of individual tumours, based on the catalogue of somatic mutations they have accumulated[3,14,15]. These inferences include timing of chromosomal gains during early somatic evolution[16], phylogenetic analysis of late cancer evolution using matched primary and metastatic tumour samples from individual patients[17–20], and temporal ordering of driver mutations across many samples[21,22].

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [2]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [3]Wellcome Sanger Institute, Cambridge, UK. [4]The Francis Crick Institute, London, UK. [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Big Data Institute, University of Oxford, Oxford, UK. [7]University of Cambridge, Cambridge, UK. [8]University of Toronto, Toronto, Ontario, Canada. [9]Vector Institute, Toronto, Ontario, Canada. [10]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [11]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [12]German Cancer Research Center (DKFZ), Heidelberg, Germany. [13]Heidelberg University, Heidelberg, Germany. [14]University of Ljubljana, Ljubljana, Slovenia. [15]NorthShore University HealthSystem, Evanston, IL, USA. [16]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [17]Simon Fraser University, Burnaby, British Columbia, Canada. [18]Vancouver Prostate Centre, Vancouver, British Columbia, Canada. [19]University of Melbourne, Melbourne, Victoria, Australia. [20]Walter and Eliza Hall Institute, Melbourne, Victoria, Australia. [21]University of Leuven, Leuven, Belgium. [22]Weill Cornell Medicine, New York, NY, USA. [23]New York Genome Center, New York, NY, USA. [24]University of California Santa Cruz, Santa Cruz, CA, USA. [25]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [26]University of California, Los Angeles, CA, USA. [27]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [28]Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. [29]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [30]Harvard Medical School, Boston, MA, USA. [31]Dana-Farber Cancer Institute, Boston, MA, USA. [32]Indiana University, Bloomington, IN, USA. [33]The University of Chicago, Chicago, IL, USA. [34]University of Cologne, Cologne, Germany. [35]University of Helsinki, Helsinki, Finland. [36]University of Glasgow, Glasgow, UK. [37]A list of members and their affiliations appears at the end of the paper. [38]Oxford NIHR Biomedical Research Centre, Oxford, UK. [39]A list of members and their affiliations appears online. [920]These authors contributed equally: Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez. [921]These authors jointly supervised this work: Paul T. Spellman, David C. Wedge, Peter Van Loo. *e-mail: moritz.gerstung@ebi.ac.uk; peter.vanloo@crick.ac.uk

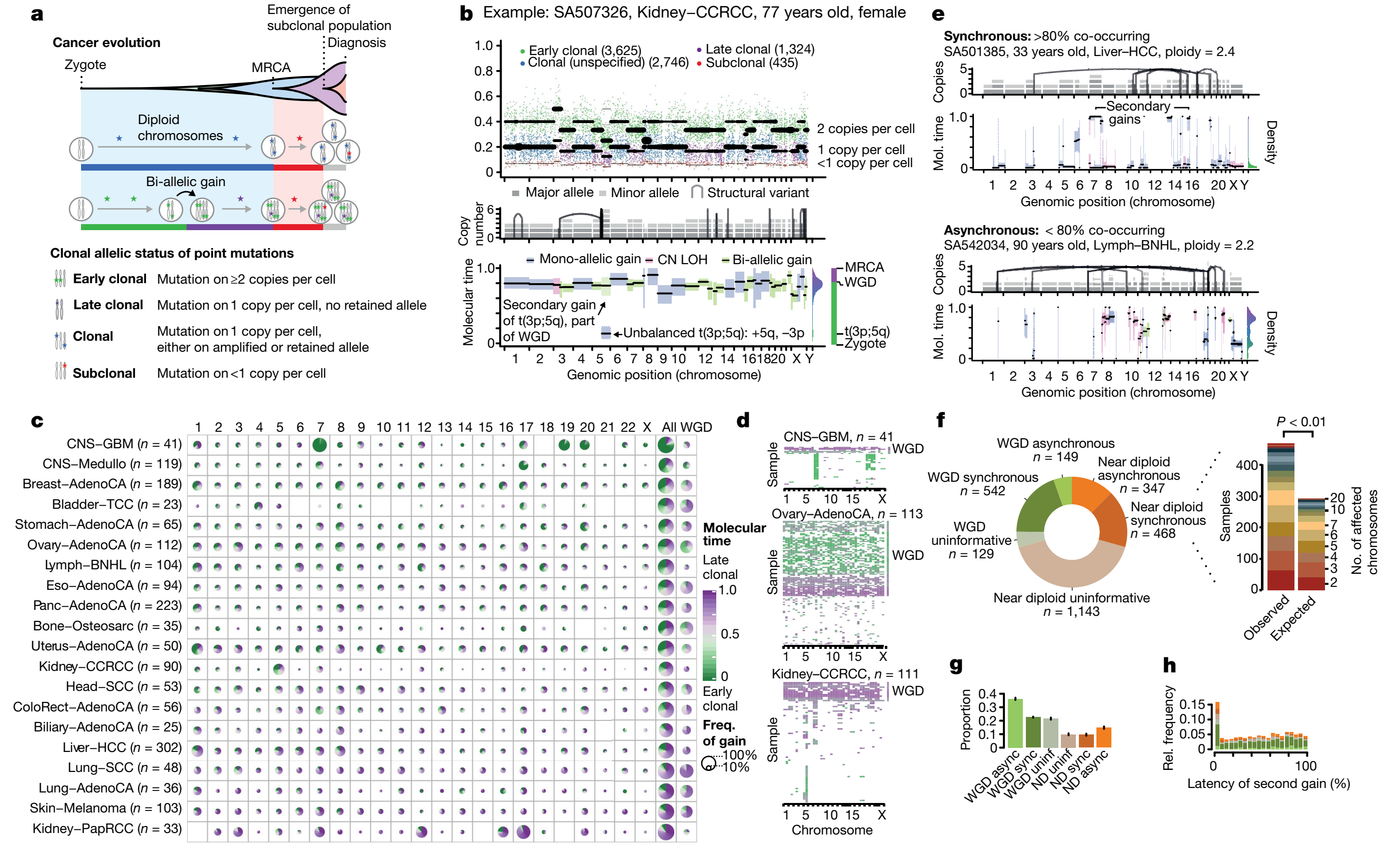


Fig. 1 | Timing clonal copy number gains using allele frequencies of point mutations. a, Principles of timing mutations and copy number gains based on whole-genome sequencing. The number of sequencing reads reporting point mutations can be used to discriminate variants as early or late clonal (green or purple, respectively) in cases of specific copy number gains, as well as clonal (blue) or subclonal (red) in cases without. **b**, Annotated point mutations in one sample based on VAF (top), copy number (CN) state and structural variants (middle), and resulting timing estimates (bottom). LOH, loss of heterozygosity. **c**, Overview of the molecular timing distribution of copy number gains across cancer types. Pie charts depict the distribution of the inferred mutation time for a given copy number gain in a cancer type. Green denotes early clonal gains, with a gradient to purple for late gains. The size of each chart is proportional to the recurrence of this event. Abbreviations for each cancer type are defined in Supplementary Table 1. **d**, Heat maps representing molecular timing estimates of gains on different chromosome arms (*x* axis) for individual samples (*y* axis) for selected tumour types. **e**, Temporal patterns of two near-diploid cases illustrating synchronous gains (top) and asynchronous gains (bottom). **f**, Left, distribution of synchronous and asynchronous gain patterns across samples, split by WGD status. Uninformative samples have too few or too small gains for accurate timing. Right, the enrichment of synchronous gains in near-diploid samples is shown by systematic permutation tests. **g**, Proportion of copy number segments ($n = 90{,}387$) with secondary gains. Error bars denote 95% credible intervals. ND, near diploid. **h**, Distribution of the relative latency of $n = 824$ secondary gains with available timing information, scaled to the time after the first gain and aggregated per chromosome.

The PCAWG Consortium has aggregated whole-genome sequencing data from 2,658 cancers[4], generated by the ICGC and TCGA, and produced high-accuracy somatic variant calls, driver mutations, and mutational signatures[4,23,24] (Methods and Supplementary Information).

Here, we leverage the PCAWG dataset to characterize the evolutionary history of 2,778 cancer samples from 2,658 unique donors across 38 cancer types. We infer timing and patterns of chromosomal evolution and learn typical sequences of mutations across samples of each cancer type. We then define broad periods of tumour evolution and examine how drivers and mutational signatures vary between these epochs. Using clock-like mutational processes, we map mutation timing estimates into approximate real time. Combined, these analyses allow us to sketch out the typical evolutionary trajectories of cancer, and map them in real time relative to the point of diagnosis.

Reconstructing the life history of tumours

The genome of a cancer cell is shaped by the cumulative somatic aberrations that have arisen during its evolutionary past, and part of this history can be reconstructed from whole-genome sequencing data[3] (Fig. 1a). Initially, each point mutation occurs on a single chromosome in a single cell, which gives rise to a lineage of cells bearing the same mutation. If that chromosomal locus is subsequently duplicated, any point mutation on this allele preceding the gain will subsequently be present on the two resulting allelic copies, unlike mutations succeeding the gain, or mutations on the other allele. As sequencing data enable the measurement of the number of allelic copies, one can define categories of early and late clonal variants, preceding or succeeding copy number gains, as well as unspecified clonal variants, which are common to all cancer cells, but cannot be timed further. Lastly, we identify subclonal mutations, which are present in only a subset of cells and have occurred after the most recent common ancestor (MRCA) of all cancer cells in the tumour sample (Supplementary Information).

The ratio of duplicated to non-duplicated mutations within a gained region can be used to estimate the time point when the gain happened during clonal evolution, referred to here as molecular time, which measures the time of occurrence relative to the total number of (clonal) mutations. For example, there would be few, if any, co-amplified early clonal mutations if the gain had occurred right after fertilization, whereas a gain that happened towards the end of clonal tumour evolution would contain many duplicated mutations[14] (Fig. 1a, Methods).

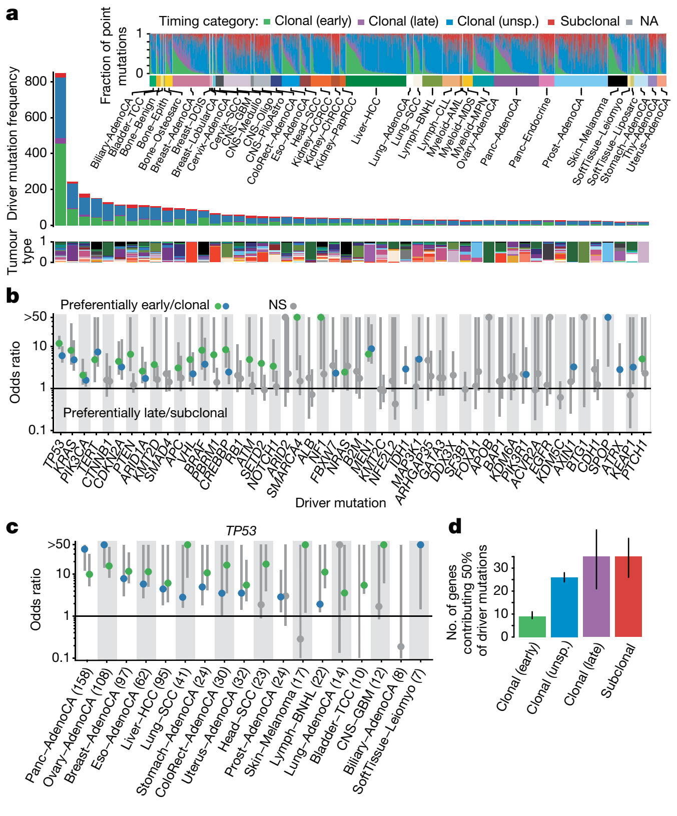


Fig. 2 | Timing of point mutations shows that recurrent driver gene mutations occur early. a, Top, distribution of point mutations over different mutation periods in $n = 2,778$ samples. Middle, timing distribution of driver mutations in the 50 most recurrent lesions across $n = 2,583$ white listed samples from unique donors. Bottom, distribution of driver mutations across cancer types; colour as defined in the inset. **b**, Relative timing of the 50 most recurrent driver lesions, calculated as the odds ratio of early versus late clonal driver mutations versus background, or clonal versus subclonal. Error bars denote 95% confidence intervals derived from bootstrap resampling. Odds ratios overlapping 1 in less than 5% of bootstrap samples are considered significant (coloured). The underlying number of samples with a given mutation is shown in **a**. **c**, Relative timing of *TP53* mutations across cancer types, as in **b**. The number of samples is defined in the x-axis labels. **d**, Estimated number of unique lesions (genes) contributing 50% of all driver mutations in different timing epochs across $n = 2,583$ unique samples, containing $n = 5,756$ driver mutations with available timing information. Error bars denote the range between 0 and 1 pseudocounts; bars denote the average of the two values. NA, not applicable; NS, not significant.

These analyses are illustrated in Fig. 1b. As expected, the variant allele frequencies (VAFs) of somatic point mutations cluster around the values imposed by the purity of the sample, local copy number configuration and identified subclonal populations. The depicted clear cell renal cell carcinoma has gained chromosome arm 5q at an early molecular time as part of an unbalanced translocation t(3p;5q), which confirms the notion that this lesion often occurs in adolescence in this cancer type[16]. At a later time point, the sample underwent a whole genome duplication (WGD) event, duplicating all alleles, including the derivative chromosome, in a single event, as evidenced by the mutation time estimates of all copy number gains clustering around a single time point, independently of the exact copy number state.

Timing patterns of copy number gains

To systematically examine the mutational timing of chromosomal gains throughout the evolution of tumours in the PCAWG dataset, we applied this analysis to the 2,116 samples with copy number gains suitable for timing (Supplementary Information). We find that chromosomal gains occur across a wide range of molecular times (median molecular time 0.60, interquartile range (IQR) 0.10–0.87), with systematic differences between tumour types, whereas within tumour types, different chromosomes typically show similar distributions (Fig. 1c, Extended Data Figs. 1, 2, Supplementary Information). In glioblastoma and medulloblastoma, a substantial fraction of gains occurs early in molecular time. By contrast, in lung cancers, melanomas and papillary kidney cancers, gains arise towards the end of the molecular timescale. Most tumour types, including breast, ovarian and colorectal cancers, show relatively broad periods of chromosomal instability, indicating a very variable timing of gains across samples.

There are, however, certain tumour types with consistently early or late gains of specific chromosomal regions. Most pronounced is glioblastoma, in which 90% of tumours contain single copy gains of chromosome 7, 19 or 20 (Fig. 1c, d). Notably, these gains are consistently timed within the first 10% of molecular time, which suggests that they arise very early in a patient's lifetime. In the case of trisomy 7, typically less than 3 out of 600 single nucleotide variants (SNVs) on the whole chromosome precede the gain (Extended Data Fig. 3a, b). On the basis of a mutation rate of $\mu = 4.8 \times 10^{-10}$ to 3.0×10^{-9} SNVs per base pair per division[25], this indicates that the trisomy occurs within the first 6–39 cell divisions, suggesting a possible early developmental origin, in agreement with somatic mosaicisms observed in the healthy brain[26]. Similarly, the duplications leading to isochromosome 17q in medulloblastoma are timed exceptionally early (Extended Data Fig. 3c, d).

Notably, we observed that gains in the same tumour often appear to occur at a similar molecular time, pointing towards punctuated bursts of copy number gains involving most gained segments (Fig. 1e). Although this is expected in tumours with WGD (Fig. 1b), it may seem surprising to observe synchronous gains in near-diploid tumours, particularly as only 6% of co-amplified chromosomal segments were linked by a direct inter-chromosomal structural variant. Still, synchronous gains are frequent, occurring in 57% (468 out of 815) of informative near-diploid tumours, 61% more frequently than expected by chance ($P < 0.01$, permutation test; Fig. 1f). Because most arm-level gains increment the allele-specific copy number by 1 (80–90%; Fig. 1g), it seems that these gains arise through mis-segregation of single copies during anaphase. This notion is further supported by the observation that in about 85% of segments with two gains of the same allele, the second gain appears with noticeable latency after the first (Fig. 1h). Therefore, the extensive chromosome-scale copy number aberrations observed in many cancer genomes are seemingly caused by a limited number of events—possibly by merotelic attachments of chromosomes to multipolar mitotic spindles[27], or as a consequence of negative selection of individual aneuploidies[28]—offering an explanation for observations of punctuated evolution in breast and colorectal cancer[29,30].

Timing of point mutations in driver genes

As outlined above, point mutations (SNVs and insertions and deletions (indels)) can be qualitatively assigned to different epochs, allowing the timing of driver mutations. Out of the 47 million point mutations in 2,583 unique samples, 22% were early clonal, 7% late clonal, 53% unspecified clonal and 17% subclonal (Fig. 2a). Among a panel of 453 cancer driver genes, 5,913 oncogenic point mutations were identified[4], of which 29% were early clonal, 5% late clonal, 56% unspecified clonal and 8% subclonal. It thus emerges that common drivers are enriched in the early clonal and unspecified clonal categories and depleted in the late clonal and subclonal ones, indicating a preferential early timing (Fig. 2b). For example, driver mutations in *TP53* and *KRAS* are 12 and 8 times enriched in early clonal stages, respectively. For *TP53*, this trend is independent of tumour type (Fig. 2c). Mutations in *PIK3CA* are two times more frequently clonal than expected, and non-coding changes near the *TERT* gene are three times more frequently early clonal.

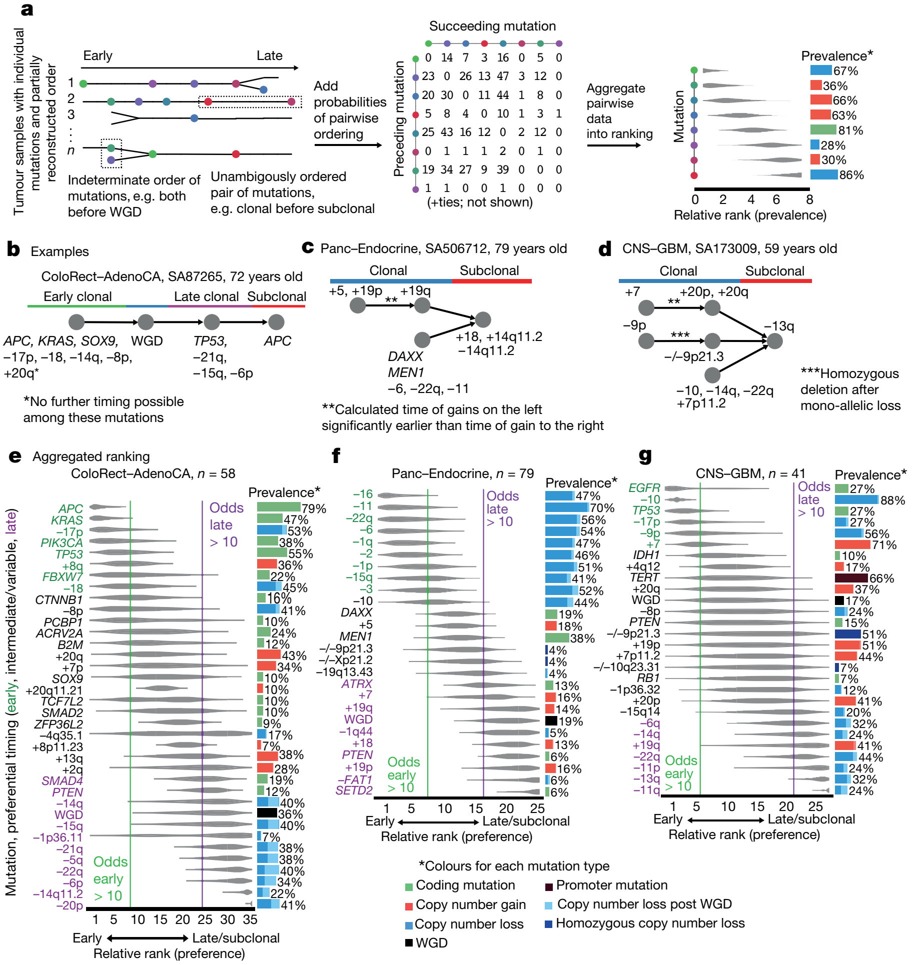


Fig. 3 | Aggregating single-sample ordering reveals typical timing of driver mutations. a, Schematic representation of the ordering process. **b**–**d**, Examples of individual patient trajectories (partial ordering relationships), the constituent data for the ordering model process. **e**–**g**, Preferential ordering diagrams for colorectal adenocarcinoma (ColoRect–AdenoCA) (**e**), pancreatic neuroendocrine cancer (Panc–Endocrine) (**f**) and glioblastoma (CNS–GBM) (**g**). Probability distributions show the uncertainty of timing for specific events in the cohort. Events with odds above 10 (either earlier or later) are highlighted. The prevalence of the event type in the cohort is displayed as a bar plot on the right.

Aggregating the clonal status of all driver point mutations over time reveals an increased diversity of driver genes mutated at later stages of tumour development: 50% of all early clonal driver mutations occur in just 9 genes, whereas 50% of late and subclonal mutations occur in approximately 35 different genes each, a nearly fourfold increase (Fig. 2d). Consistent with previous studies of individual tumour types[31–34], these results suggest that, in general, the very early events in cancer evolution occur in a constrained set of common drivers, and a more diverse array of drivers is involved in late tumour development.

Relative timing of somatic driver events

Although timing estimates of individual events reflect evolutionary periods that differ from one sample to another, they define in part the order in which driver mutations and copy number alterations have occurred in each sample (Fig. 3a–d). As confirmed by simulations, aggregating these orderings across samples defines a probabilistic ranking of lesions (Fig. 3a), recapitulating whether each mutation occurs preferentially early or late during tumour evolution (Extended Data Figs. 4, 5, Supplementary Information).

In colorectal adenocarcinoma, for example, we find *APC* mutations to have the highest odds of occurring early, followed by *KRAS*, loss of 17p and *TP53*, and *SMAD4* (Fig. 3b, e). Whole-genome duplications occur after tumours have accumulated several driver mutations, and many chromosomal gains and losses are typically late. These results are in agreement with the classical *APC-KRAS-TP53* progression model of Fearon and Vogelstein[35], but add considerable detail.

In many cancer types, the sequence of events during cancer progression has not previously been determined in detail. For example, in pancreatic neuroendocrine cancers, we find that many chromosomal losses, including those of chromosomes 2, 6, 11 and 16, are among the earliest events, followed by driver mutations in *MEN1* and *DAXX* (Fig. 3c, f). WGD events occur later, after many of these tumours have reached a pseudo-haploid state due to widespread chromosomal losses. In glioblastoma, we find that the loss of chromosome 10, and driver mutations in *TP53* and *EGFR* are very early, often preceding early gains of chromosomes 7, 19 and 20 (Fig. 3d, g). Mutations in the *TERT* promoter tend to occur at early to intermediate time points, whereas other driver mutations and copy number changes tend to be later events.

Across cancer types, we typically find *TP53* mutations among the earliest events, as well as losses of chromosome 17 (Supplementary Information). WGD events usually have an intermediate ranking, and most copy number changes occur later. Losses typically precede gains, and consistent with the results above, common drivers typically occur before rare drivers.

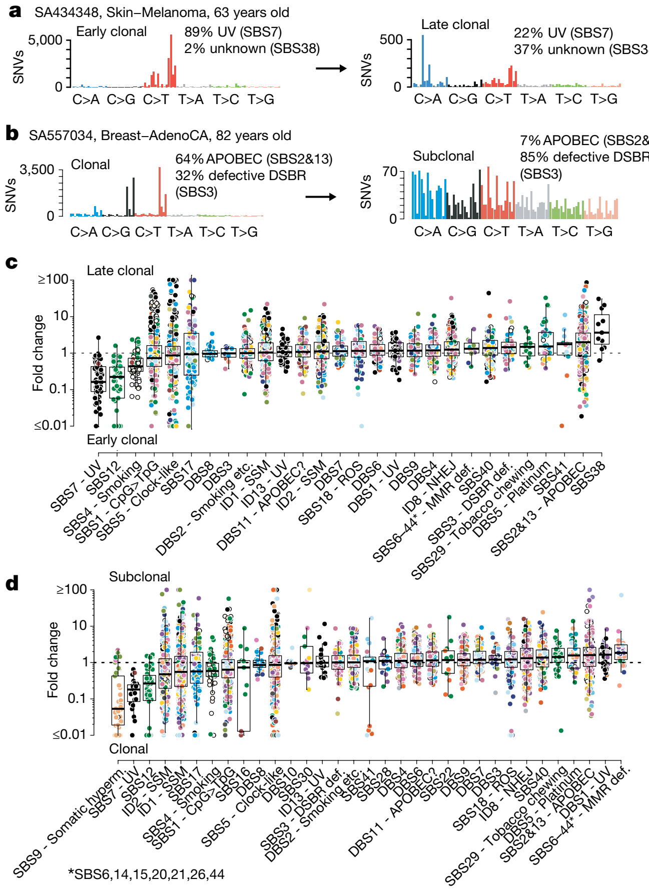


Fig. 4 | Dynamic mutational processes during early and late clonal tumour evolution. a, Example of tumours with substantial changes between mutation spectra of early (left) and late (right) clonal time points. The attribution of mutations to the most characteristic signatures are shown. **b**, Example of clonal-to-subclonal mutation spectrum change. **c**, Fold changes between relative proportions of early and late clonal mutations attributed to individual mutational signatures. Points are coloured by tissue type. Data are shown for samples ($n = 530$) with measurable changes in their overall mutation spectra and restricted to signatures active in at least 10 samples. Box plots demarcate the first and third quartiles of the distribution, with the median shown in the centre and whiskers covering data within 1.5× the IQR from the box. **d**, Fold changes between clonal and subclonal periods in samples ($n = 729$) with measurable changes in their mutation spectra, analogous to **c**.

Timing of mutational signatures

The cancer genome is shaped by various mutational processes over its lifetime, stemming from exogenous and cell-intrinsic DNA damage, and error-prone DNA replication, leaving behind characteristic mutational spectra, termed mutational signatures[24,36]. Stratifying mutations by their clonal allelic status, we find evidence for a changing mutational spectrum between early and late clonal time points in 29% (530 out of 1,852) of informative samples ($P < 0.05$, Bonferroni-adjusted likelihood-ratio test), typically changing the spectrum by 19% (median absolute difference; range 4–66%) (Fig. 4a, b, Extended Data Fig. 6). Similarly, 30% of informative samples (729 out of 2,387) displayed changes of their mutation spectrum between the clonal and subclonal state, with median difference of 21% (range 3–72%). Combined, the mutation spectrum changes throughout tumour evolution in 40% of samples (1,069 out of 2,688).

To quantify whether the observed temporal changes can be attributed to known and suspected mutational processes, we decomposed the mutational spectra at each time point into a catalogue of 57 mutational signatures, including double base substitution and indel signatures[24] (Methods).

In general, these mutational signatures display a predominantly undirected temporal variability over several orders of magnitude (Fig. 4c, d, Extended Data Fig. 7). In addition, several signatures demonstrate distinct temporal trends. As one may expect, signatures of exogenous mutagens are predominantly active in the early clonal stages of tumorigenesis. These include tobacco smoking in lung adenocarcinoma (signature SBS4, median fold change 0.43, IQR 0.31–0.72), consistent with previous reports[37,38], and ultraviolet light exposure in melanoma (SBS7; median fold change 0.16, IQR 0.09–0.43). Another strong decrease over time is found for a signature of unknown aetiology, SBS12, which acts mostly in liver cancers (median fold change 0.22, IQR 0.06–0.41). In chronic lymphoid leukaemia, there was a 20-fold relative decrease in mutations associated with somatic hypermutation (SBS9; median fold change 0.05, IQR 0.02–0.43) from clonal to subclonal stages.

Some mutational processes tend to increase throughout cancer evolution. For example, we see that APOBEC mutagenesis (SBS2 and SBS13) increases in many cancer types from the early to late clonal stages (median fold change 2.0, IQR 0.8–3.6), as does a newly described signature SBS38 (median fold 3.6, IQR 1.8–11). Signatures of defective mismatch repair (SBS6, 14, 15, 20, 21, 26 and 44) increase from clonal to subclonal stages (median fold 1.8, IQR 1.2–3.0).

Chronological time estimates

The molecular timing data presented above do not measure the occurrence of events in chronological time. If the rate at which mutations are acquired per year in each sample was constant, the chronological time would simply be the product of the estimated molecular timing and age at diagnosis. However, this relation will be nonlinear if the mutation rate changes over time, and is inflated by acquired mutational processes, as suggested by the analysis in the previous section. Some of these issues can be mitigated by counting only mutations contributed by endogenous and less variable mutational processes, such as CpG-to-TpG mutations (hereafter CpG>TpG) caused by spontaneous deamination of 5-methyl-cytosine to thymine at CpG dinucleotides, which have been proposed as a molecular clock[12]. Our supplementary analysis suggests that, although the baseline CpG>TpG mutation rate in cancers is very close to that in normal cells, there appears to be a moderate increase (1–10 times, adding between 20 and 40% of mutations) in cancers (Extended Data Fig. 8). As this shifts chronological timing estimates, we model different scenarios of the evolution of the CpG>TpG mutation rate (Fig. 5a).

Applying this logic to time WGDs, which yield sufficient numbers of CpG>TpG mutations, demonstrates that they occur several years and possibly even a decade or more before diagnosis in some cancer types, under a range of scenarios of mutation rate increase (Fig. 5b, Extended Data Fig. 9). A notable example is ovarian adenocarcinoma, which appears to have a median latency of more than 10 years. This holds true even under a scenario of a CpG>TpG rate increase of 20-fold, which would be far beyond the 7.5-fold rate increase observed in matched primary and relapse samples[39] (Extended Data Fig. 8f). Notably, these results suggest WGD may occur throughout the entire female reproductive life (Extended Data Fig. 9b). The latency between the MRCA and the last detectable subclone is shorter, typically several months to years (Fig. 5c).

These timescales of cancer evolution are further supported by the fact that progression of most known precancerous lesions to carcinomas usually spans many years, if not decades[40–45]. Our data corroborate these timescales and extend them to cancer types without detectable premalignant conditions, raising the hope that these tumours could also be detected in less malignant stages.

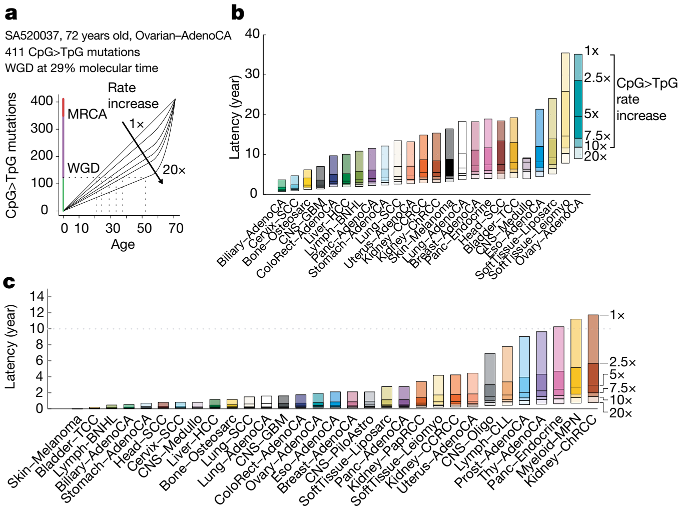


Fig. 5 | Approximate chronological timing inference suggests a timescale of cancer evolution of several years. a, Mapping of molecular timing estimates to chronological time under different scenarios of increases in the CpG>TpG mutation rate. A greater increase before diagnosis indicates an inflation of the mutation timescale. **b**, Median latency between WGDs and the last detectable subclone before diagnosis under different scenarios of CpG>TpG mutation rate increases for $n = 569$ non-hypermutant cancers with at least 100 informative SNVs, low tumour in normal contamination and at least five samples per tumour histology. **c**, Median latency between the MRCA and the last detectable subclone before diagnosis for different CpG>TpG mutation rate changes in $n = 1{,}921$ non-hypermutant samples with low tumour in normal contamination and at least 5 cases per cancer type.

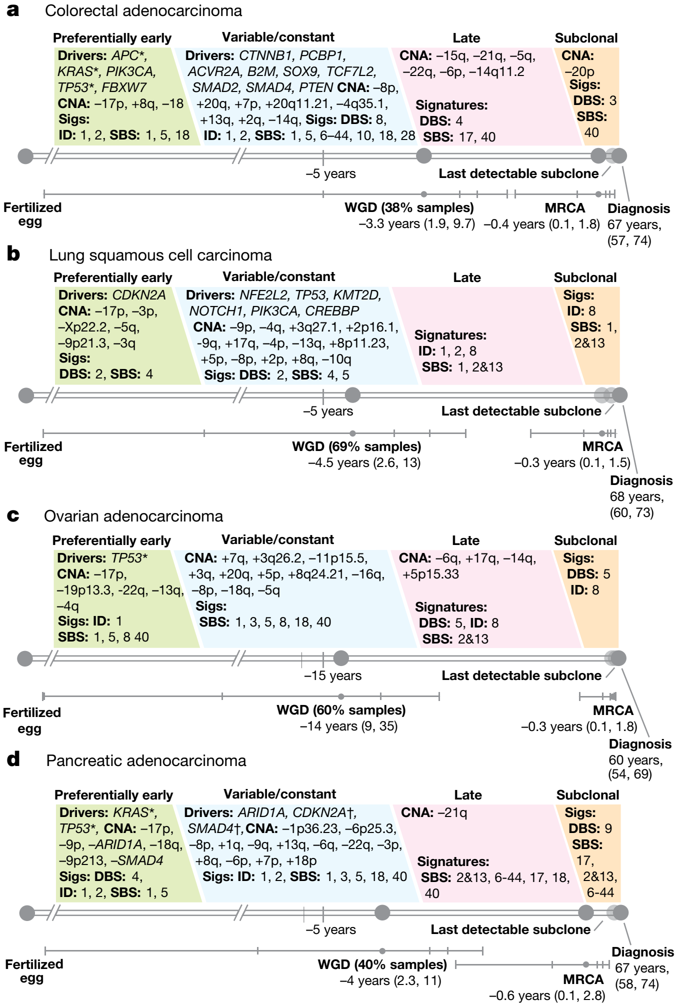

a
Colorectal adenocarcinoma
Preferentially early
Drivers: APC*, KRAS*, PIK3CA, TP53*, FBXW7
CNA: –17p, +8q, –18
Sigs: ID: 1, 2, SBS: 1, 5, 18
Variable/constant
Drivers: CTNNB1, PCBP1, ACVR2A, B2M, SOX9, TCF7L2, SMAD2, SMAD4, PTEN CNA: –8p, +20q, +7p, +20q11.21, –4q35.1, +13q, +2q, –14q, Sigs: DBS: 8, ID: 1, 2, SBS: 1, 5, 6–44, 10, 18, 28
Late
CNA: –15q, –21q, –5q, –22q, –6p, –14q11.2
Signatures: DBS: 4 SBS: 17, 40
Subclonal
CNA: –20p
Sigs: DBS: 3 SBS: 40
–5 years
Last detectable subclone
Fertilized egg
WGD (38% samples)
–3.3 years (1.9, 9.7)
MRCA
–0.4 years (0.1, 1.8)
Diagnosis
67 years, (57, 74)
b
Lung squamous cell carcinoma
Preferentially early
Drivers: CDKN2A
CNA: –17p, –3p, –Xp22.2, –5q, –9p21.3, –3q
Sigs: DBS: 2, SBS: 4
Variable/constant
Drivers: NFE2L2, TP53, KMT2D, NOTCH1, PIK3CA, CREBBP
CNA: –9p, –4q, +3q27.1, +2p16.1, -9q, +17q, –4p, –13q, +8p11.23, +5p, –8p, +2p, +8q, –10q
Sigs: DBS: 2, SBS: 4, 5
Late
Signatures: ID: 1, 2, 8 SBS: 1, 2&13
Subclonal
Sigs: ID: 8 SBS: 1, 2&13
–5 years
Last detectable subclone
Fertilized egg
WGD (69% samples)
–4.5 years (2.6, 13)
MRCA
–0.3 years (0.1, 1.5)
Diagnosis
68 years, (60, 73)
c
Ovarian adenocarcinoma
Preferentially early
Drivers: TP53*
CNA: –17p, –19p13.3, -22q, –13q, –4q
Sigs: ID: 1
SBS: 1, 5, 8 40
Variable/constant
CNA: +7q, +3q26.2, –11p15.5, +3q, +20q, +5p, +8q24.21, –16q, –8p, –18q, –5q
Sigs: SBS: 1, 3, 5, 8, 18, 40
Late
CNA: –6q, +17q, –14q, +5p15.33
Signatures: DBS: 5, ID: 8 SBS: 2&13
Subclonal
Sigs: DBS: 5 ID: 8
–15 years
Last detectable subclone
Fertilized egg
WGD (60% samples)
–14 years (9, 35)
MRCA
–0.3 years (0.1, 1.8)
Diagnosis
60 years, (54, 69)
d
Pancreatic adenocarcinoma
Preferentially early
Drivers: KRAS*, TP53*, CNA: –17p, –9p, –ARID1A, –18q, –9p213, –SMAD4
Sigs: DBS: 4, ID: 1, 2, SBS: 1, 5
Variable/constant
Drivers: ARID1A, CDKN2A†, SMAD4†, CNA: –1p36.23, –6p25.3, –8p, +1q, –9q, +13q, –6q, –22q, –3p, +8q, –6p, +7p, +18p
Sigs: ID: 1, 2, SBS: 1, 3, 5, 18, 40
Late
CNA: –21q
Signatures: SBS: 2&13, 6-44, 17, 18, 40
Subclonal
Sigs: DBS: 9 SBS: 17, 2&13, 6-44
–5 years
Last detectable subclone
Fertilized egg
WGD (40% samples)
–4 years (2.3, 11)
MRCA
–0.6 years (0.1, 2.8)
Diagnosis
67 years, (58, 74)

Fig. 6 | Typical timelines of tumour development. a–**d**, Timelines representing the length of time, in years, between the fertilized egg and the median age of diagnosis for colorectal adenocarcinoma (**a**), squamous cell lung cancer (**b**), ovarian adenocarcinoma (**c**) and pancreatic adenocarcinoma (**d**). Real-time estimates for major events, such as WGD and the emergence of the MRCA, are used to define early, variable, late and subclonal stages of tumour evolution approximately in chronological time. The range of chronological time estimates according to varying clock mutation acceleration rates is shown as well, with tick marks corresponding to 1×, 2.5×, 5×, 7.5×, 10× and 20×. Driver mutations and copy number alterations (CNA) are shown in each stage according to their preferential timing, as defined by relative ordering. Mutational signatures (Sigs) that, on average, change over the course of tumour evolution, or are substantially active but not changing, are shown in the epoch in which their activity is greatest. DBS, double base substitution; SBS, single base substitutions. Where applicable, lesions with a known timing from the literature are annotated; dagger symbols denotes events that were found to have a different timing; asterisk symbol denotes events that agree with our timing.

Discussion

To our knowledge, our study presents the first large-scale genome-wide reconstruction of the evolutionary history of cancers, reconstructing both early (pre-cancer) and later stages of 38 cancer types. This is facilitated by the timing of copy number gains relative to all other events in the genome, through multiplicity and clonal status of co-amplified point mutations. However, several limitations exist (Supplementary Information). Perhaps most importantly, molecular timing is based on point mutations and is therefore subject to changes in mutation rate. Notably, healthy tissues acquire point mutations at rates not too dissimilar from those seen in cancers, particularly when considering only endogenous mutational processes, and furthermore, some tissues are riddled with microscopic clonal expansions of driver gene mutations[5–9,11]. This is direct evidence that the life history of almost every cell in the human body, including those that develop into cancer, is driven by somatic evolution.

Together, the data presented here enable us to draw approximate timelines summarizing the typical evolutionary history of each cancer type (Fig. 6, Supplementary Information for all other cancer types). These make use of the qualitative timing of point mutations and copy number alterations, as well as signature activities, which can be interleaved with the chronological estimates of WGD and the appearance of the MRCA.

It is remarkable that the evolution of practically all cancers displays some level of order, which agrees very well with, and adds much detail to, established models of cancer progression[35,46]. For example, *TP53* with accompanying 17p deletion is one of the most frequent initiating mutations in a variety of cancers, including ovarian cancer, in which it is the hallmark of its precancerous precursor lesions[47]. Furthermore, the list of typically early drivers includes most other highly recurrent cancer genes, such as *KRAS*, *TERT* and *CDKN2A*, indicating a preferred role in early and possibly even pre-cancer evolution. This initially constrained set of genes broadens at later stages of cancer development, suggesting an epistatic fitness landscape canalizing the first steps of cancer evolution. Over time, as tumours evolve, they follow increasingly diverse paths driven by individually rare driver mutations, and by copy number alternations. However, none of these trends is absolute, and the evolutionary paths of individual tumours are highly variable, showing that cancer evolution follows trends, but is far from deterministic.

Our study sheds light on the typical timescales of in vivo tumour development, with initial driver events seemingly occurring up to decades before diagnosis, demonstrating how cancer genomes are shaped by a lifelong process of somatic evolution, with fluid boundaries between normal ageing processes[5–11] and cancer evolution.

Nevertheless, the presence of genetic aberrations with such long latency raises hopes that aberrant clones could be detected early, before reaching their full malignant potential.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1907-7.

1. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).
2. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
3. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
4. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
5. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Preprint at bioRxiv https://doi.org/10.1101/505685 (2018).
6. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
7. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
8. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
9. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
10. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
11. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
12. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet*. **47**, 1402–1407 (2015).
13. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
14. Durinck, S. et al. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov*. **1**, 137–143 (2011).
15. Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol*. **19**, 95 (2018).
16. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623 (2018).
17. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med*. **366**, 883–892 (2012).
18. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
19. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med*. **21**, 751–759 (2015).
20. Brastianos, P. K. et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov*. **5**, 1164–1177 (2015).
21. Papaemmanuil, E. et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627 (2013).
22. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
23. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1965-x (2020).
24. Alexandrov, L. B. The repertoire of mutational signatures in human cancer. *Nature* https://doi.org/10.1038/s41586-020-1943-3 (2020).
25. Keogh, M. J. et al. High prevalence of focal and multi-focal somatic genetic variants in the human brain. *Nat. Commun*. **9**, 4257 (2018).
26. Heim, S. et al. Trisomy 7 and sex chromosome loss in human brain tissue. *Cytogenet. Cell Genet*. **52**, 136–138 (1989).
27. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278–282 (2009).
28. Sheltzer, J. M. et al. Single-chromosome gains commonly function as tumor suppressors. *Cancer Cell* **31**, 240–255 (2017).
29. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet*. **48**, 1119–1130 (2016).
30. Cross, W. et al. The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol*. **2**, 1661–1672 (2018).
31. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet*. **46**, 225–233 (2014).
32. Gibson, W. J. et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat. Genet*. **48**, 848–855 (2016).
33. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017).
34. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med*. **376**, 2109–2121 (2017).
35. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
36. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
37. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med*. **7**, 283ra54 (2015).
38. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*. **17**, 31 (2016).
39. Patch, A.-M. et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
40. Bostwick, D. G. & Qian, J. High-grade prostatic intraepithelial neoplasia. *Mod. Pathol*. **17**, 360–379 (2004).
41. Brenner, H. et al. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut* **56**, 1585–1589 (2007).
42. Gazdar, A. F. & Brambilla, E. Preneoplasia of lung cancer. *Cancer Biomark*. **9**, 385–396 (2010).
43. Sanders, M. E., Schuyler, P. A., Dupont, W. D. & Page, D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* **103**, 2481–2484 (2005).
44. Schlecht, N. F. et al. Human papillomavirus infection and time to progression and regression of cervical intraepithelial neoplasia. *J. Natl. Cancer Inst*. **95**, 1336–1343 (2003).
45. Whitson, M. J. & Falk, G. W. Predictors of progression to high-grade dysplasia or adenocarcinoma in Barrett's esophagus. *Gastroenterol. Clin. North Am*. **44**, 299–315 (2015).
46. Bardeesy, N. & DePinho, R. A. Pancreatic cancer biology and genetics. *Nat. Rev. Cancer* **2**, 897–909 (2002).
47. Folkins, A. K. et al. A candidate precursor to pelvic serous cancer (p53 signature) and its prevalence in ovaries and fallopian tubes from women with BRCA mutations. *Gynecol. Oncol*. **109**, 168–173 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PCAWG Evolution & Heterogeneity Working Group

Stefan C. Dentro[3,4,6], Ignaty Leshchiner[5], Moritz Gerstung[1,2,3], Clemency Jolly[4], Kerstin Haase[4], Maxime Tarabichi[3,4], Jeff Wintersinger[8,9], Amit G. Deshwar[8,9], Kaixian Yu[11], Santiago Gonzalez[1], Yulia Rubanova[8,9], Geoff Macintyre[16], David J. Adams[3], Pavana Anur[10], Rameen Beroukhim[5,31], Paul C. Boutros[8,25,26], David D. Bowtell[27], Peter J. Campbell[3], Shaolong Cao[11], Elizabeth L. Christie[19,27], Marek Cmero[19,20], Yupeng Cun[34], Kevin J. Dawson[3], Jonas Demeulemeester[4,21], Nilgun Donmez[17,18], Ruben M. Drews[16], Roland Eils[12,13], Yu Fan[11], Matthew Fittall[4], Dale W. Garsed[19,27], Gad Getz[5,28,29,30], Gavin Ha[5], Marcin Imielinski[22,23], Lara Jerman[1,14], Yuan Ji[15,33], Kortine Kleinheinz[12,13], Juhee Lee[24], Henry Lee-Six[3], Dimitri G. Livitz[5], Salem Malikic[17,18], Florian Markowetz[16], Inigo Martincorena[3], Thomas J. Mitchell[3,7], Ville Mustonen[35], Layla Oesper[40], Martin Peifer[34], Myron Peto[10], Benjamin J. Raphael[41], Daniel Rosebrock[5], S. Cenk Sahinalp[18,32], Adriana Salcedo[25], Matthias Schlesner[12], Steven Schumacher[5], Subhajit Sengupta[15], Ruian Shi[8], Seung Jun Shin[11,42], Oliver Spiro[5], Lincoln D. Stein[25], Ignacio Vázquez-García[3,7], Shankar Vembu[8], David A. Wheeler[43], Tsun-Po Yang[34], Xiaotong Yao[22,23], Ke Yuan[16,36], Hongtu Zhu[11], Wenyi Wang[11], Quaid D. Morris[8,9], Paul T. Spellman[10], David C. Wedge[6,38] & Peter Van Loo[4,21]

[40]Department of Computer Science, Carleton College, Northfield, MN, USA. [41]Department of Computer Science, Princeton University, Princeton, NJ, USA. [42]Korea University, Seoul, South Korea. [43]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

Methods

Dataset

The PCAWG series consists of 2,778 tumour samples (2,703 white listed, 75 grey listed) from 2,658 donors. All samples in this dataset underwent whole-genome sequencing (minimum average coverage 30× in the tumour, 25× in the matched normal samples), and were processed with a set of project-specific pipelines for alignment, variant calling, and quality control[4]. Copy number calls were established by combining the output of six individual callers into a consensus using a multitier approach, resulting in a copy number profile, a purity and ploidy value and whether the tumour has undergone a WGD (Supplementary Information). Consensus subclonal architectures have been obtained by integrating the output of 11 subclonal reconstruction callers, after which all SNVs, indels and structural variants are assigned to a mutation cluster using the MutationTimer.R approach (Supplementary Information). Driver calls have been defined by the PCAWG Driver Working Group[4], and mutational signatures are defined by the PCAWG Signatures Working Group[24]. A more detailed description can be found in Supplementary Information, section 1.

Data accrual was based on sequencing experiments performed by individual member groups of the ICGC and TCGA, as described in an associated study[4]. As this is a meta-analysis of existing data, power calculations were not performed and the investigators were not blinded to cancer diagnoses.

Timing of gains

We used three related approaches to calculate the timing of copy number gains (see Supplementary Information, section 2). In brief, the common feature is that the expected VAF of a mutation (E) is related to the underlying number of alleles carrying a mutation according to the formula: $E[X] = nmf\rho/[N(1-\rho) + C\rho]$, in which X is the number of reads, n denotes the coverage of the locus, the mutation copy number m is the number of alleles carrying the mutation (which is usually inferred), f is the frequency of the clone carrying the given mutation ($f = 1$ for clonal mutations). N is the normal copy number (2 on autosomes, 1 or 2 for chromosome X and 0 or 1 for chromosome Y), C is the total copy number of the tumour, and ρ is the purity of the sample.

The number of mutations n_m at each allelic copy number m then informs about the time when the gain has occurred. The basic formulae for timing each gain are, depending on the copy number configuration:

$$\text{Copy number } 2+1: T = 3n_2/(2n_2 + n_1)$$

$$\text{Copy number } 2+2: T = 2n_2/(2n_2 + n_1)$$

$$\text{Copy number } 2+0: T = 2n_2/(2n_2 + n_1)$$

in which 2 + 1 refers to major and minor copy number of 2 and 1, respectively. Methods differ slightly in how the number of mutations present on each allele are calculated and how uncertainty is handled (Supplementary Information).

Timing of mutations

The mutation copy number m and the clonal frequency f is calculated according to the principles indicated above. Details can be found in Supplementary Information, section 2. Mutations with $f = 1$ are denoted as 'clonal', and mutations with $f < 1$ as 'subclonal'. Mutations with $f = 1$ and $m > 1$ are denoted as 'early clonal' (co-amplified). In cases with $f = 1$, $m = 1$ and $C > 2$, mutations were annotated as 'late clonal', if the minor copy number was 0, otherwise 'clonal' (unspecified).

Timing of driver mutations

A catalogue of driver point mutations (SNVs and indels) was provided by the PCAWG Drivers and Functional Interpretation Group[4]. The timing category was calculated as above. From the four timing categories, the odds ratios of early/late clonal and clonal (early, late or unspecified clonal)/subclonal were calculated for driver mutations against the distribution of all other mutations present in fragments with the same copy number composition in the samples with each particular driver. The background distribution of these odds ratios was assessed with 1,000 bootstraps (Supplementary Information, section 4.1).

Integrative timing

For each pair of driver point mutations and recurrent copy number alterations, an ordering was established (earlier, later or unspecified). The information underlying this decision was derived from the timing of each driver point mutation, as well as from the timing status of clonal and subclonal copy number segments. These tables were aggregated across all samples and a sports statistics model was employed to calculate the overall ranking of driver mutations. A full description is given in Supplementary Information, section 4.2.

Timing of mutational signatures

Mutational trinucleotide substitution signatures, as defined by the PCAWG Mutational Signatures Working Group[24], were fit to samples with observed signature activity, after splitting point mutations into either of the four epochs. A likelihood ratio test based on the multinomial distribution was used to test for differences in the mutation spectra between time points. Time-resolved exposures were calculated using non-negative linear least squares. Full details are given in Supplementary Information, section 5.

Real-time estimation of WGD and MRCA

CpG>TpG mutations were counted in an NpCpG context, except for skin–melanoma, in which CpCpG and TpCpG were excluded owing to the overlapping UV mutation spectrum. For visual comparison, the number of mutations was scaled to the effective genome size, defined as the 1/mean(m_i/C_i), in which m_i is the estimated number of allelic copies of each mutation, and C_i is the total copy number at that locus, thereby scaling to the final copy number and the time of change.

A hierarchical Bayesian linear regression was fit to relate the age at diagnosis to the scaled number of mutations, ensuring positive slope and intercept through a shared gamma distribution across cancer types.

For tumours with several time points, the set of mutations shared between diagnosis and relapse (n_D) and those specific to the relapse (n_R) was calculated. The rate acceleration was calculated as: $a = n_R/n_D \times t_D/t_R$. This analysis was performed separately for all substitutions and for CpG>TpG mutations.

On the basis of these analyses, a typical increase of 5× for most cancer types was chosen, with a lower value of 2.5× for brain cancers and a value of 7.5× for ovarian cancer.

The correction for transforming an estimate of a copy number gain in mutation time into chronological time depends not only on the rate acceleration, but also on the time at which this acceleration occurred. As this is generally unknown, we performed Monte Carlo simulations of rate accelerations spanning an interval of 15 years before diagnosis, corresponding roughly to 25% of time for a diagnosis at 60 years of age, noting that a 5× rate increase over this duration yields an offset of about 33% of mutations, compatible with our data. Subclonal mutations were assumed to occur at full acceleration. The proportion of subclonal mutations was divided by the number of identified subclones, thus conservatively assuming branching evolution. Full details are given in Supplementary Information, section 6.

Cancer timelines

The results from each of the different timing analyses are combined in timelines of cancer evolution for each tumour type (Fig. 6 and Supplementary Information). Each timeline begins at the fertilized egg, and spans up to the median age of diagnosis within each cohort. Real-time estimates for WGD and the MRCA act as anchor points, allowing us to roughly map the four broadly defined time periods (early clonal, intermediate, late clonal and subclonal) to chronological time during a patient's lifespan. Specific driver mutations or copy number alterations can be placed within each of these time frames based on their ordering from the league model analysis. Signatures are shown if they typically change over time (95% confidence intervals of mean change not overlapping 0), and if they are strongly active (contributing at least 10% mutations to one time point). Signatures are shown on the timeline in the epoch of their greatest activity. Where an event found in our study has a known timing in the literature, the agreement is annotated on the timeline; with an asterisk denoting an agreed timing, and dagger symbol denoting a timing that is different to our results. Full details are given in Supplementary Information, section 7.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA PCAWG Consortium are described elsewhere[4] and available for download at https://dcc.icgc.org/releases/PCAWG. Further information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic SNVs derived from TCGA donors, researchers will also need to obtain dbGaP authorization. Datasets used and results presented in this study, including timing estimates for copy number gains, chronological estimates of WGD and MRCA, as well as mutation signature changes, are described in Supplementary Note 3 and are available at https://dcc.icgc.org/releases/PCAWG/evolution-heterogeneity.

Code availability

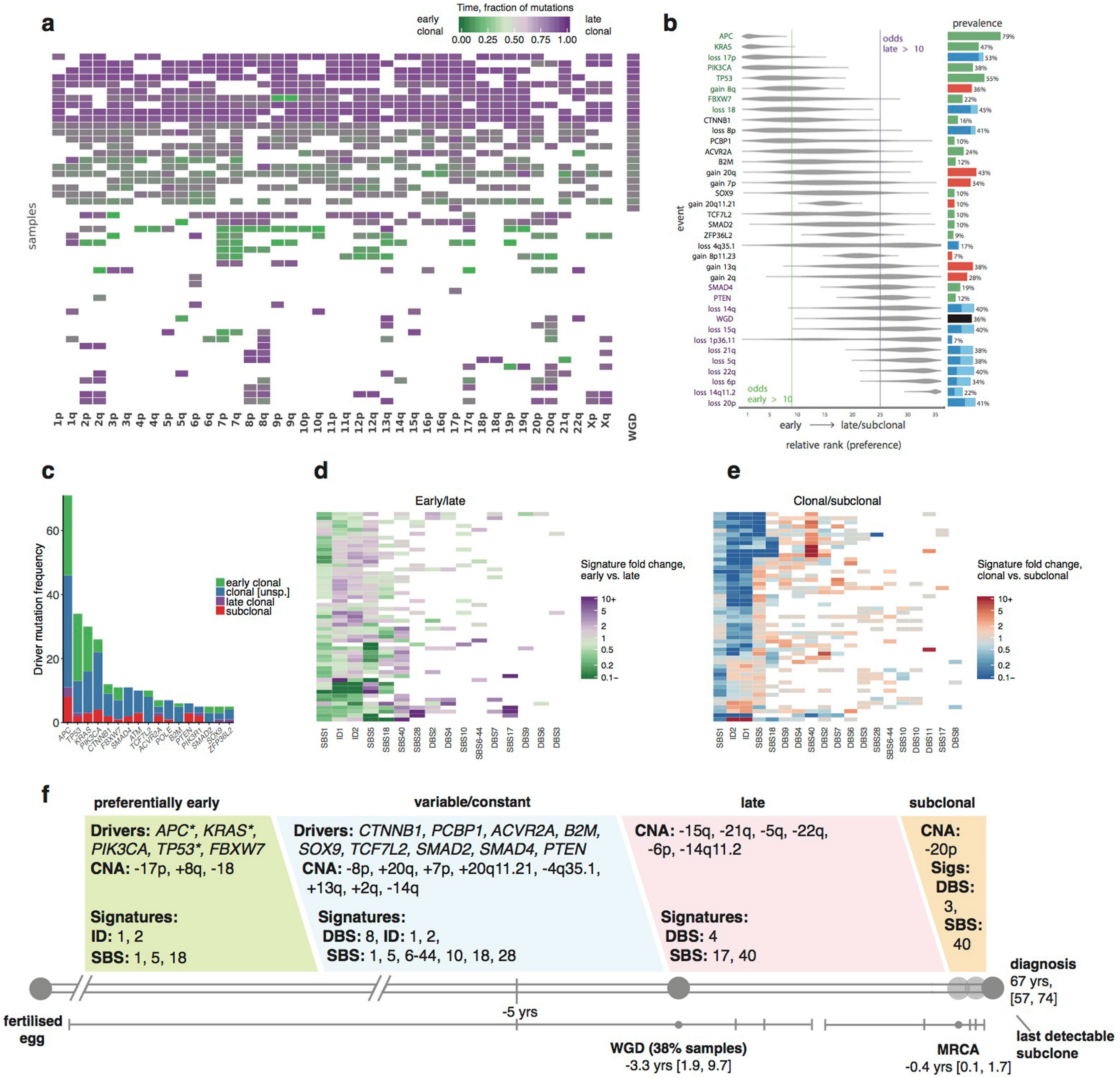
The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v3.0, which allows for reuse and distribution. Analysis code presented in this study is available through the GitHub repository https://github.com/PCAWG-11/Evolution. This archive contains relevant software and analysis workflows as submodules, which include code for timing copy number gains, point mutations and mutation signatures, real-time timing and evolutionary league model analysis, as well as scripts to generate the figures presented: CancerTiming (v.3.1.8), MutationTimeR (v.0.1), PhylogicNDT (v.1.1) and a series of custom scripts (v. 1.0), with detailed versions of other packages used.

Acknowledgements We thank H. Lee-Six and L. Moore for sharing data on mutation burden in normal tissues. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). This project was enabled through the Crick Scientific Computing STP and through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (grant number MR/L016311/1). M.T. and J.D. are postdoctoral fellows supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie grant agreement number 747852-SIOMICS and 703594-DECODE). J.D. is a postdoctoral fellow of the FWO. F.M., G.M. and K. Yuan acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. G.M., K. Yuan and F.M. were funded by CRUK core grants C14303/A17197 and A19274. S. Sengupta and Y.J. are supported by NIH R01 CA132897. S.M. is supported by the Vanier Canada Graduate Scholarship. S.C.S. is supported by the NSERC Discovery Frontiers Project, "The Cancer Genome Collaboratory" and NIH Grant GM108308. H.Z. is supported by grant NIMH086633 and an endowed Bao-Shan Jing Professorship in Diagnostic Imaging. W.W. is supported by the US National Cancer Institute (1R01 CA183793 and P30 CA016672). P.T.S. was supported by U24CA210957 and 1U24CA143799. D.C.W. is funded by the Li Ka Shing foundation. P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

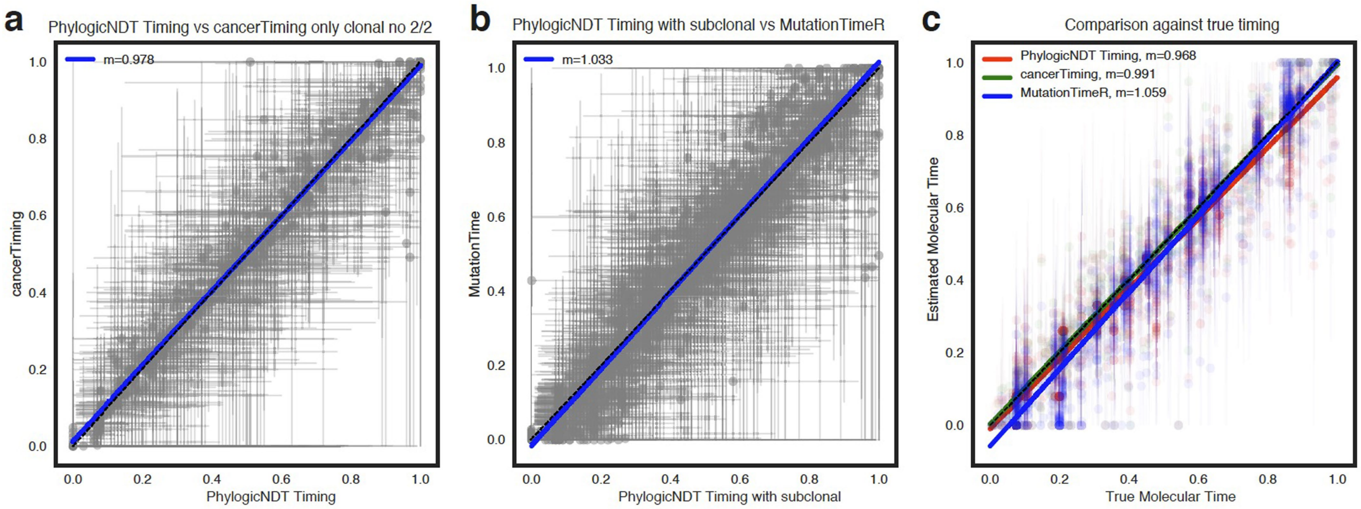
Author contributions M.G., C.J., I.L., S.G., P.A., D.R., D.G.L., P.T.S. and P.V.L. performed timing of point mutations and copy number gains. S.G. and M.G. performed qualitative timing of driver point mutations and analyses of synchronous gains, L.J. timed secondary copy number gains. I.L., T.J.M., D.R., D.G.L., D.C.W. and G.G. performed relative timing of somatic driver events and implemented integrative models. C.J., Y.R., M.G., Q.D.M. and P.V.L. performed timing of mutational signatures. M.G. performed real-time estimation of whole-genome duplication and subclonal diversification. S.G. assessed mutation rates in relapsed samples. C.J., M.G., I.L., Y.R., D.R. and P.V.L. constructed cancer timelines. M.G., C.J., I.L., S.C.D., S.G., T.J.M., Y.R., P.A., J.D., P.C.B., D.D.B., V.M., Q.D.M., P.T.S., D.C.W. and P.V.L. interpreted the results. S.C.D., I.L., J.W., A.D., I.V.-G., K. Yuan, G.M., M.P., S.M., N.D., K. Yu, S. Sengupta, K.H., M.T., J.D., D.G.L., D.R., J.L., M.C., S.C.S., Y.J., F.M., V.M., H.Z., W.W., Q.D.M., D.C.W. and P.V.L. performed subclonal architecture analysis. S.C.D., I.L., K.K., V.M., M.P., X.Y., D.G.L., S. Schumacher, R.B., M.I., M.S., D.C.W. and P.V.L. performed copy number analysis. J.W., S.C.D., I.L., K.H., D.G.L., K.K., D.R., D.C.W., Q.D.M. and P.V.L. derived a consensus of copy number analysis results. K. Yu, M.T., A.D., S.C.D., I.L., D.C.W., M.G., P.V.L., Q.D.M. and W.W. derived a consensus of subclonal architecture results. Y.F. and W.W. contributed to subclonal mutation calls. P.T.S., D.C.W. and P.V.L. coordinated the study. M.G., C.J., P.T.S., Y.R., I.L., Q.D.M., D.C.W. and P.V.L. wrote the manuscript, which all authors approved. S.C.D., I.L., M.G., C.J., K.H., M.T., J.W., A.G.D., K. Yu, S.G., Y.R. and G.M. in the PCAWG Evolution & Heterogeneity Working Group contributed equally. W.W., Q.D.M., P.T.S., D.C.W. and P.V.L. in the PCAWG Evolution & Heterogeneity Working Group jointly supervised the work.

Competing interests R.B. owns equity in Ampressa Therapeutics. G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect and POLYSOLVER. I.L. is a consultant for PACT Pharma. B.J.R. is a consultant at and has ownership interest (including stock and patents) in Medley Genomics. All other authors declare no competing interests.

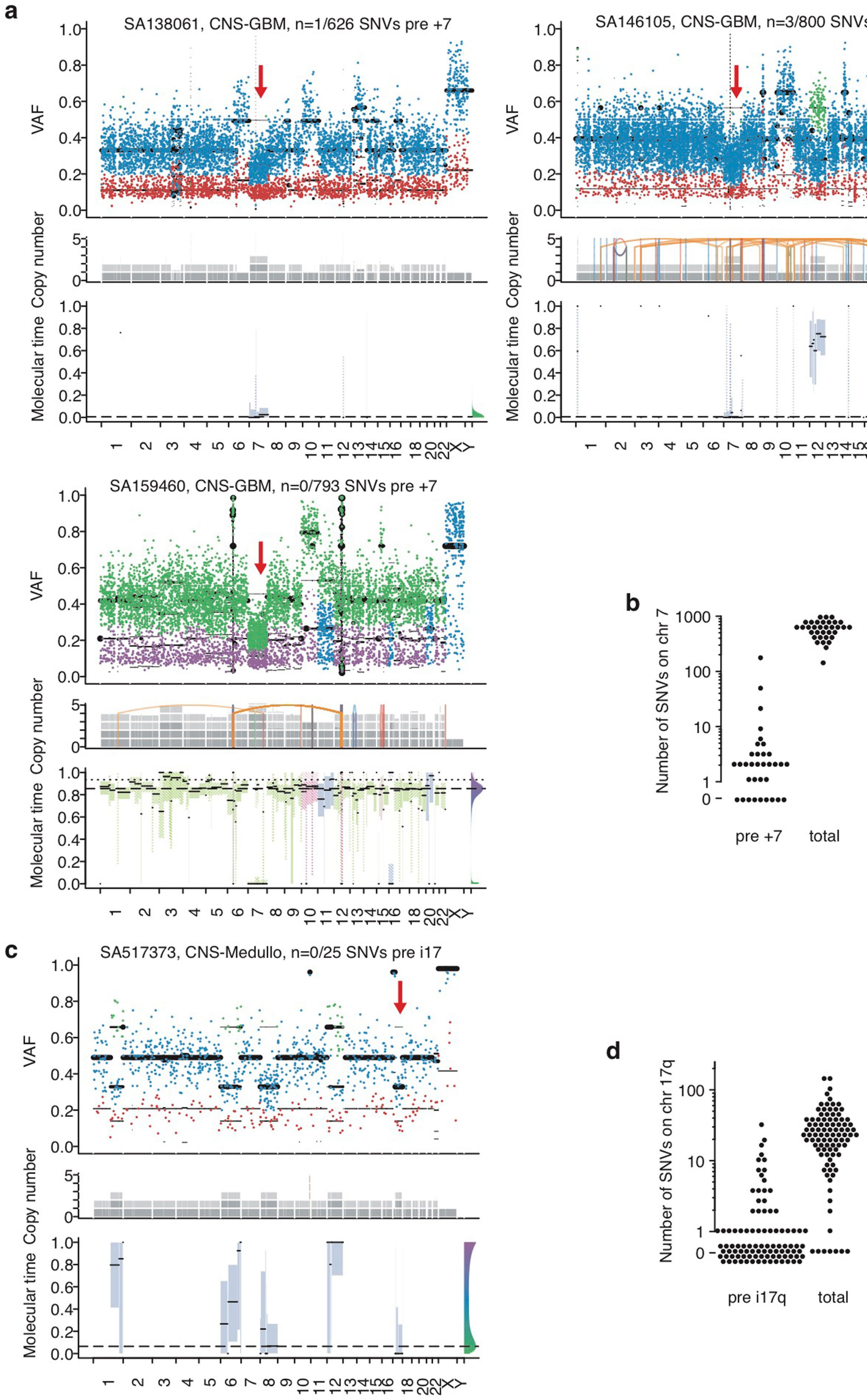
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1907-7.
Correspondence and requests for materials should be addressed to M.G. or P.V.L.
Reprints and permissions information is available at http://www.nature.com/reprints.



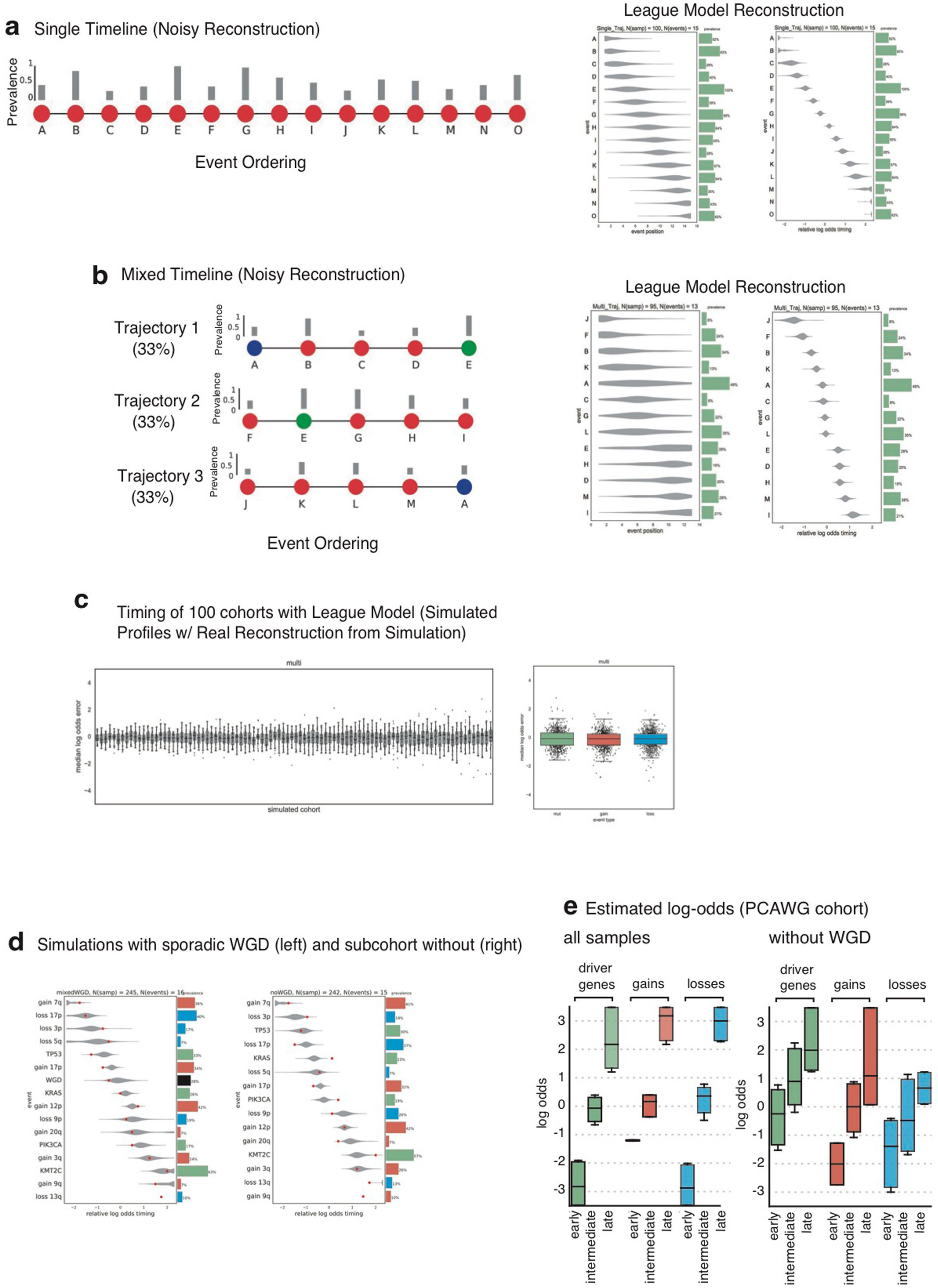
Extended Data Fig. 1 | Summary of all results obtained for colorectal adenocarcinoma ($n = 60$) as an example. a, Clustered heat maps of mutational timing estimates for gained segments, per patient. Colours as indicated in main text: green represents early clonal events, purple represents late clonal. **b**, Relative ordering of copy number events and driver mutations across all samples. **c**, Distribution of mutations across early clonal, late clonal and subclonal stages, for the most common driver genes. A maximum of 10 driver genes are shown. **d**, Clustered mutational signature fold changes between early clonal and late clonal stages, per patient. Green and purple indicate, respectively, a signature decrease and increase in late clonal from early clonal mutations. Inactive signatures are coloured white. **e**, As in **d** but for clonal versus subclonal stages. Blue indicates a signature decrease and red an increase in subclonal from clonal mutations. **f**, Typical timeline of tumour development. Similar result summaries for all other cancer types can be found in the Supplementary Information (pages 46–77).



Extended Data Fig. 2 | Comparison of methods used for timing of individual copy number gains. a, **b**, Pairwise comparison of the three approaches for timing individual copy number gains. **c**, Comparison using simulated data, showing high concordance.

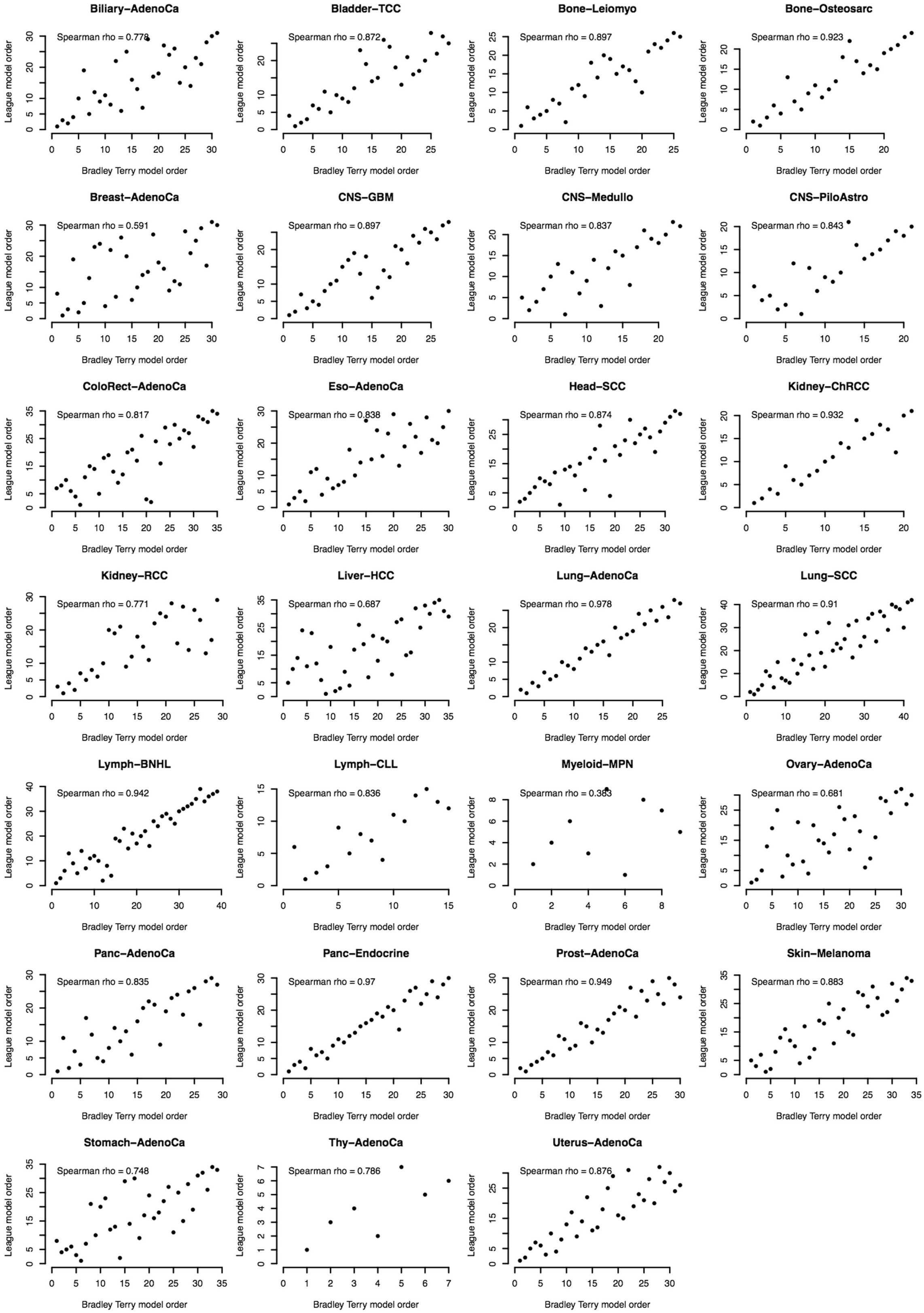


Extended Data Fig. 3 | Early copy number gains in brain cancers. a, Three illustrative examples of glioblastoma with trisomy 7. The red arrow depicts the expected VAF cluster of point mutations preceding trisomy 7, which usually contains less than three SNVs. **b**, Distributions of the number of SNVs preceding trisomy 7 and total number of mutations on chromosome (chr) 7 in *n* = 34 GBM samples with trisomy 7. **c**, Medulloblastoma example with isochromosome 17q. **d**, Distributions of SNVs on 17q in *n* = 95 samples with isochromosome 17q; 74 out of 95 samples have less than 1 SNV preceding the isochromosome.

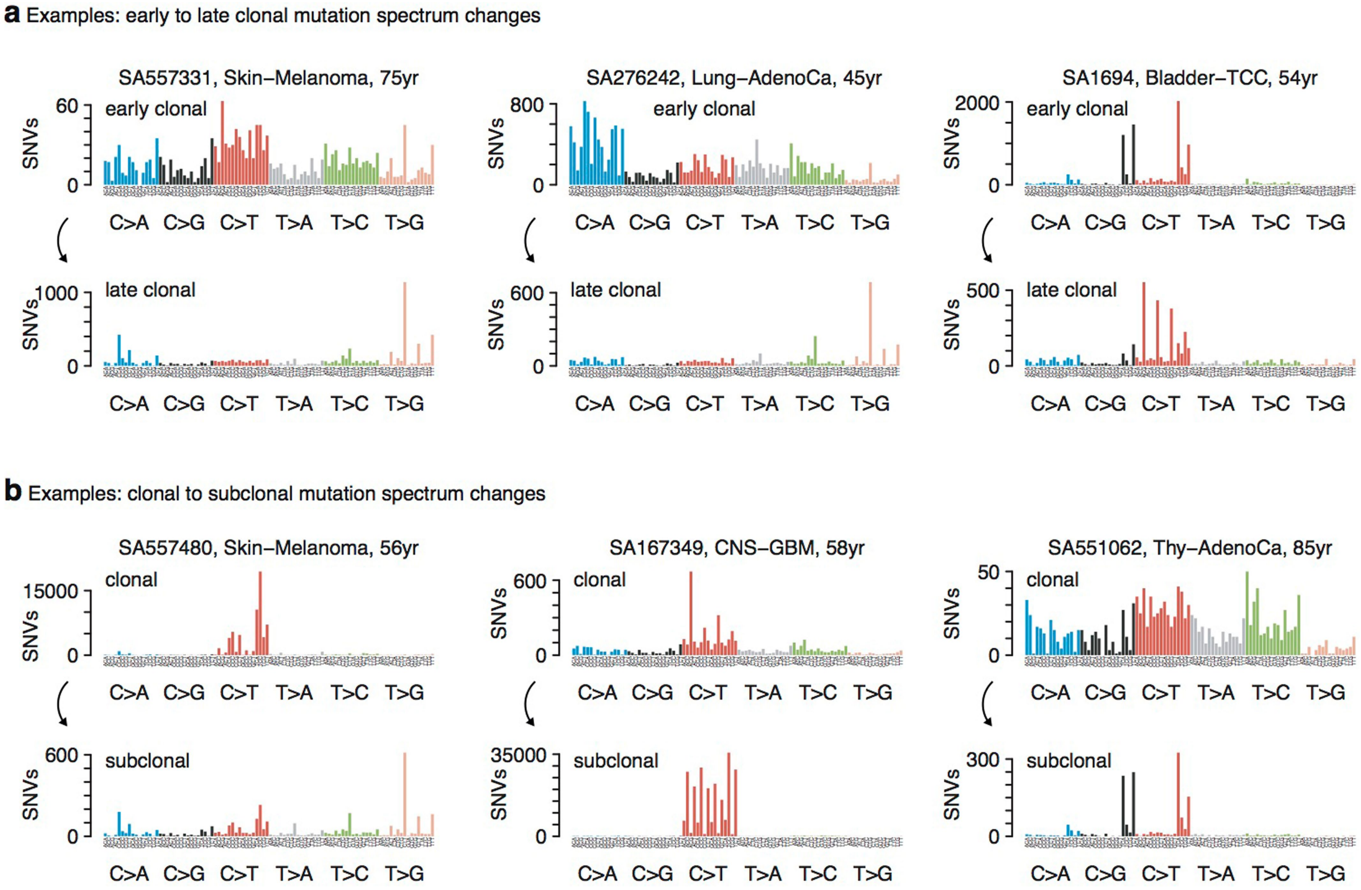


Extended Data Fig. 4 | See next page for caption.

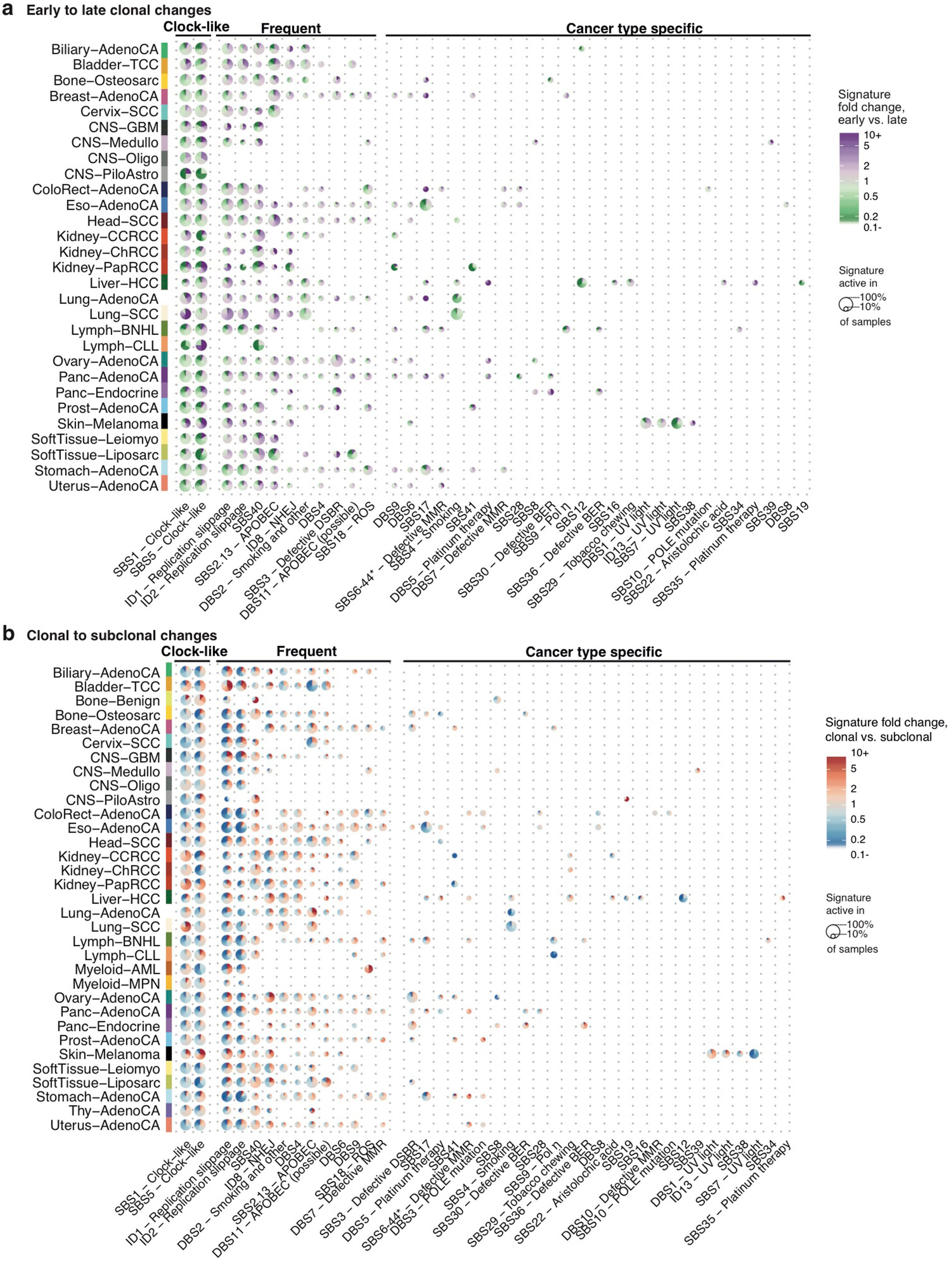
Extended Data Fig. 4 | Validation of relative ordering model reconstruction based on simulated cohorts of whole-genome samples. a, Relative ordering model (PhylogicNDT LeagueModel) results for a simulated cohort of samples (n = 100) from a single generalized relative order of events (with varied prevalence) showing high concordance with the true trajectory. Probability distributions show the uncertainty of timing for specific events in the cohort. **b**, Relative ordering model results on a simulated cohort of samples (n = 95) from a complex mixture of trajectories with different order of events showing high concordance with the expected average trajectory. **c**, Estimation of accuracy of the relative ordering model reconstruction by simulation of a set of 100 cohorts (n(samples) = 100) with random trajectory mixtures and quantifying the distance in log odds early/late from perfect ordering. For the vast majority of events (even with low number of occurrences in the cohort), the log odds error does not exceed 1, confirming that very few events would switch between timing categories. The inset box corresponds to the first and third quartiles of the distribution, the horizontal line indicates the median and whiskers include data within 1.5× the IQR from the box. **d**, Simulated data show concordant timing in cohorts with WGD (n = 245). Exclusion of samples with WGD (right, n = 242) introduces only a mild drop in accuracy, indicating that WGD is beneficial but not necessary for the reconstruction. Red dot = true rank. **e**, Estimated log odds in observed data including WGD (left, n = 245) and without (right, n = 242), across different mutation types. The inset box corresponds to the first and third quartiles of the distribution, the horizontal line indicates the median and whiskers include data within 1.5× the IQR from the box.



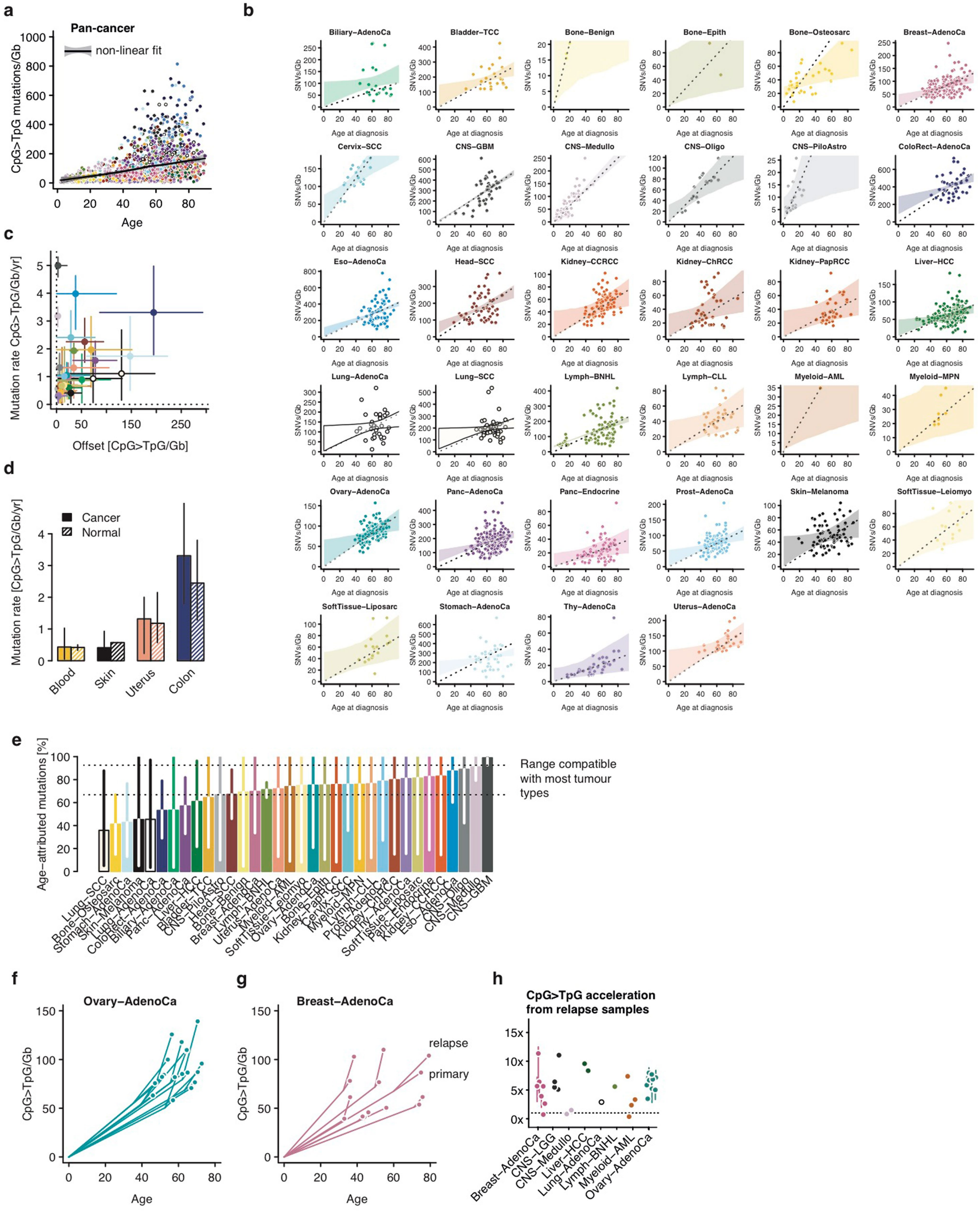
Extended Data Fig. 5 | Correlation between the league model and Bradley–Terry model ordering. Direct comparison for each tumour type of the league and Bradley–Terry models for determining the order of recurrent somatic mutations and copy number events. Axes indicate the ordered events observed in the respective tumour types. Correlation is quantified by Spearman's rank correlation coefficient. A total of *n* = 756 ordered events are shown.



Extended Data Fig. 6 | Examples of mutation spectrum changes across tumour evolution. a, Three examples of tumours with substantial changes between mutation spectra of early (top) and late (bottom) clonal time points. **b**, Three examples of tumours with substantial changes between mutation spectra of clonal (top) and subclonal (bottom) time points.

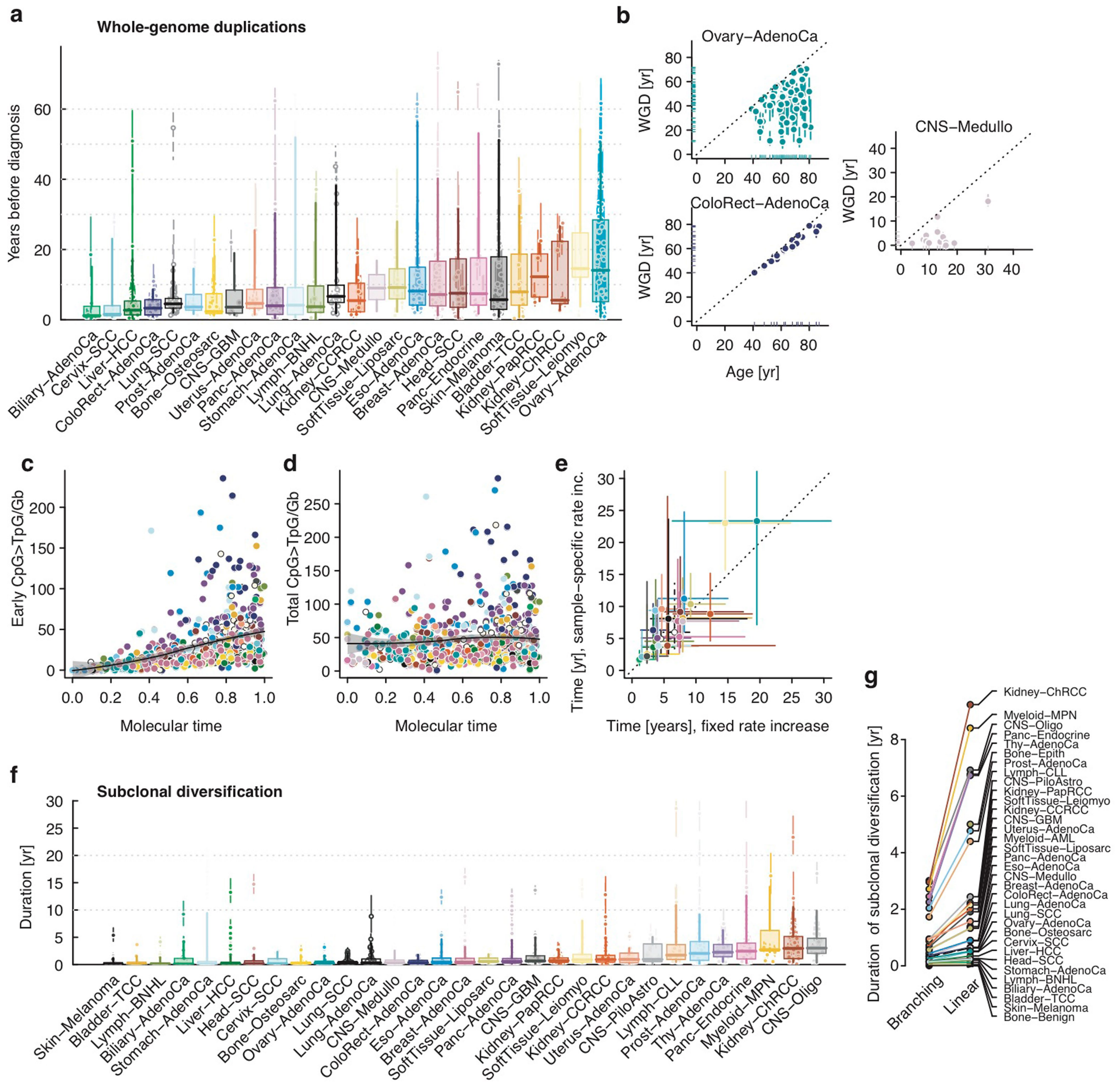


Extended Data Fig. 7 | Overview of early-to-late clonal and clonal-to-subclonal signature changes across tumour types. a, **b**, Pie charts representing signature changes per cancer type for early-to-late clonal signature changes (**a**) and clonal-to-subclonal signature changes (**b**). Signatures that decrease between early and late are coloured green; signatures that increase are purple. The size of each pie chart represents the frequency of each signature. Signatures are split into three categories: (1) clock-like, comprising the putative clock signatures 1 and 5; (2) frequent, which are signatures present in ten or more cancer types; and (3) cancer-type specific, which are in fewer than ten cancer types and are often limited to specific cohorts.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Age-dependent mutation burden and relapse samples indicate near-normal CpG>TpG mutation rate in cancer, with moderate acceleration during carcinogenesis. a, Across all cancer samples, a predominantly linear accumulation of CpG>TpG mutations (scaled to copy number) is observed over time, as measured by the age at diagnosis. **b**, Cancer-specific analysis of the CpG>TpG mutation burden as a function of age at diagnosis for *n* = 1,978 samples of 34 informative cancer types. The dotted line denotes the median mutations per year (that is, not offset), and shading denotes the 95% credible interval of a hierarchical Bayesian linear regression model across all data points. Slope and intercepts are drawn for each cancer type from a gamma distribution, respectively; inference was done by Hamiltonian Monte Carlo sampling. **c**, Maximum a posteriori estimates of rate and offset for 34 cancer types with 95% credible intervals as defined in **b**. **d**, Mutation rate inferred from cancer as in **b** and from selected normal tissue sequencing studies of *n* = 140 normal haematopoietic stem cells, *n* = 1 normal skin sample, *n* = 182 samples from normal endometrium, and *n* = 445 normal colonic crypts; error bars denote the 95% confidence interval. **e**, Median fraction of mutations attributed to linear age-dependent accumulation, based on estimates from **b** and the age at diagnosis for each sample. Error bars denote the 95% credible interval. **f**, **g**, CpG>TpG mutations per gigabase for ovarian cancer (**f**) and breast cancer (**g**) samples with matched primary and relapse samples. **h**, Increase in CpG>TpG mutation rate inferred from paired primary and relapse samples for six cancer types. Bars denote the range of the rate increase for different scenarios of copy number evolution, assuming ploidy changes have occurred prior (upper value) or posterior (lower value) to the branching between primary and relapse sample.



Extended Data Fig. 9 | Real-time estimates indicate long latencies for some samples caused by the absence of early mutations. a, Time of WGD for *n* = 571 individual patients, split by tumour type with an estimated mutation rate increase of 5×, except for ovary–adenocarcinoma (7.5×) and CNS (2.5×). Error bars represent 80% confidence intervals, reflecting uncertainty stemming from the number of mutations per segment and onset of the rate increase. Box plots demarcate the quartiles and median of the distribution with whiskers indicating 5% and 95% quantiles. **b**, Scatter plots showing the time of diagnosis (*x* axis) and inferred time of WGD (*y* axis) with error bars as in **a**. **c**, Scatter plot of early (co-amplified) CpG>TpG mutations (*y* axis) as a function of the mutational time estimate of WGD (*x* axis). The black line denotes a nonlinear loess fit with 95% confidence interval. Colours define the cancer type as in **a**. **d**, Total CpG>TpG mutations (*y* axis) as a function of the mutation time estimate of WGD (*x* axis). Colours and fit as in **c**. Early molecular timing is thus caused by a depletion of early CpG>TpG mutations, rather than an inflation of late CpG>TpG mutations. **e**, Estimated median WGD latency of *n* = 571 patients as in **a** for fixed (*x* axis) versus patient specific rate increases, depending on the observed CpG>TpG mutation burden, allowing for a higher (up to 10×) mutation rate increase in samples with more mutations (*y* axis). Error bars denote the IQR. **f**, Timing of subclonal diversification using CpG>TpG mutations in *n* = 1,953 individual patients. Box plots and error bars for data points as in **a**. **g**, Comparison of the median duration of subclonal diversification per cancer type assuming branching and linear phylogenies.

nature research

Corresponding author(s): Moritz Gerstung
Peter Van Loo

Last updated by author(s): Oct 8, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, seeAuthors & Referees and theEditorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at https://www.overture.bio. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to contact@overture.bio.
Data analysis	The PCAWG workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACEseq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15. Analysis code presented in this study is available through the github repository https://github.com/PCAWG-11/Evolution. This archive contains relevant software and analysis workflows as submodules, including code for timing copy number gains, point mutations and mutation signatures, real-time timing, and evolutionary league model analysis, as well as scripts to generate the figures presented.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. All results presented in this study, including timing estimates for copy number gains, real time estimates of WGD and MRCA, as well as mutation signature activities, are available at https://www.synapse.org/#!Synapse:syn14193595.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.
Data exclusions	After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine). Hypermutated and samples with normal contamination were excluded for chronological inferences in this study, as described in the Supplementary Methods.
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement. The accuracy of inferences in this study was assessed using simulations and by applying three different algorithms for the timing of copy number gains (Extended Data Figure 2), as well as two different algorithms for the temporal ordering of driver mutations (Extended Data Figure 5).
Randomization	N/A - This exploratory study did not contain a randomization step
Blinding	N/A - This exploratory study did not contain a blinded analysis

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patient-by-patient clinical data are provided in Extended Data Table 1 of the marker paper for the PCAWG consortium. Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.
Recruitment	Patients were recruited by the participating centres following local protocols.
Ethics oversight	The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.