The evolutionary history of 2,658 cancers

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

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [3]The Francis Crick Institute, London, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [5]Big Data Institute, University of Oxford, Oxford, United Kingdom; [6]University of Cambridge, Cambridge, United Kingdom; [7]University of Toronto, Toronto, Canada; [8]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA; [9]The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [10]German Cancer Research Center (DKFZ), Heidelberg, Germany; [11]Heidelberg University, Heidelberg, Germany; [12]NorthShore University HealthSystem, Evanston, IL, USA; [13]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; [14]Simon Fraser University, Vancouver, Canada; [15]University of Melbourne, Melbourne, Australia; [16]Department of Human Genetics, University of Leuven, Leuven, Belgium; [17]Weill Cornell Medicine, New York, NY, USA; [18]New York Genome Center, New York, NY, USA; [19]University of California Santa Cruz, Santa Cruz, CA, USA; [20]Ontario Institute for Cancer Research, Toronto, Canada; [21]Peter MacCallum Cancer Centre,

Melbourne, VIC 3052, Australia; [22]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [23]Indiana University, Bloomington, IN, USA; [24]The University of Chicago, Chicago, IL, USA; [25]University of Cologne, Cologne, Germany; [26]University of Helsinki, Helsinki, Finland; [27]University of Glasgow, Glasgow G12 8RZ, United Kingdom.

[#]These authors contributed equally.

[*]To whom correspondence may be addressed:

Moritz Gerstung, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, CB10 1SD, United Kingdom. Tel: +44 (0) 1223 49 4636, email: Moritz.Gerstung@ebi.ac.uk.

Peter Van Loo, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, United Kingdom. Tel: +44 (0) 20 3796 1719, e-mail: Peter.VanLoo@crick.ac.uk.

[28]A list of members of the PCAWG Evolution and Heterogeneity Working Group can be found at the end of the manuscript.

Summary

Cancer develops through a process of somatic evolution. Here, we reconstruct the evolutionary history of 2,778 tumour samples from 2,658 donors spanning 39 cancer types. Characteristic copy number gains, such as trisomy 7 in glioblastoma or isochromosome 17q in medulloblastoma, are found amongst the earliest events in tumour evolution. The early phases of oncogenesis are driven by point mutations in a restricted set of cancer genes, often including biallelic inactivation of tumour suppressors. By contrast, increased genomic instability, a more than three-fold diversification of driver genes, and an acceleration of mutational processes are features of later stages. Clock-like mutations yield estimates for whole genome duplications and subclonal diversification in chronological time. Our results suggest that driver mutations often precede diagnosis by many years, and in some cases decades. Taken together, these data reveal common and divergent trajectories of cancer evolution, pivotal for understanding tumour biology and guiding early cancer detection.

Introduction

Cancer arises through natural selection: initiated by mutations in a single cell, the accumulation of subsequent aberrations and the effects of selection over time result in the clonal expansions of cells, ultimately leading to the formation of a genomically aberrant tumour[1]. This model has been underpinned by genetic studies, starting with classical work on retinoblastoma[2] and the sequence of *APC*, *KRAS* and *TP53* mutations during colorectal adenoma to adenocarcinoma progression[3]. Establishing a particular order of mutations during the somatic evolution of cancers systematically across cancer types, however, has proven to be complicated due to small sample sizes and the stochastic nature of evolution between individuals.

Deep sequencing of bulk tumour samples makes it possible to examine the evolutionary history of individual tumours, based on the catalogue of somatic mutations they have accumulated[4]. Many studies have reconstructed the phylogenetic relationships between tumour samples and metastases from individual patients[5-8], corroborating the clonal evolution model. From single samples, the timing of chromosomal gains can be estimated using point mutations within duplicated regions[9,10]. In addition, the relative ordering of events within a tumour type can be determined by aggregating pairwise timing estimates of genomic changes (for example clonal *vs*. subclonal) across many samples using preference models[11,12]. While these approaches provide insights into tumour development, they have only been applied to a limited number of cancers.

Here, we use the Pan-Cancer Analysis of Whole Genomes (PCAWG)[13] dataset, as part of the International Cancer Genome Consortium (ICGC)[14] and The Cancer Genome Atlas (TCGA)[15] to characterise the evolutionary history of 2,778 cancers from 2,658 unique donors across 39 cancer types. We determine the order and timing of mutations in cancer development to delineate the patterns of chromosomal evolution across and within different cancer types. We then define broad periods of tumour evolution and examine how drivers and mutational signatures vary between these stages. Finally, using CpG>TpG mutations, we convert timing estimates into approximate real time, and create typical timelines of tumour evolution.

Results

Reconstructing the life history of single tumours

A cancer cell's genome is the cumulative result of the somatic aberrations that have arisen during its evolutionary past, and part of this history can be reconstructed from deep whole genome sequencing data (**Fig. 1a**)[4]. Initially, each point mutation occurs on a single chromosome in a single cell. If that chromosomal locus is subsequently duplicated, the point mutation will be co-amplified with the gained allele, which can be detected in deep sequencing data. Likewise, mutations found in a subset of tumour cells have not swept through the population, and must have occurred after most recent common ancestor (MRCA) of the tumour cells in the sequenced sample.

Mapping point mutations to the proportion of cells and chromosomes enables us to define three categories, which we term *early clonal*, *late clonal* and *subclonal,* each associated with broad epochs of tumour evolution (**Fig. 1a**). *Clonal* mutations have occurred before the occurrence of the MRCA and are common to all cancer cells. These can often be further subdivided as either *early clonal* if they occurred before copy number gains, or *late clonal* otherwise. Additionally, *subclonal* mutations are only observed in a fraction of cancer cells. Importantly, the number of early (and late) clonal mutations provides information about the timing of the underlying copy number segment. For example, there would be few, if any, coamplified early clonal mutations if the gain had occurred right after fertilisation (**Fig. 1a** and **Online Methods**)[9].

These analyses are illustrated in **Fig. 1b**. As expected, the frequency of somatic point mutations cluster tightly around the values imposed by the purity of the sample, local copy number configuration and identified subclones. As the sample pictured has undergone whole genome duplication (WGD), the mutation time estimates of all copy number segments scatter narrowly around a single time-point, independently of the exact copy number state, confirming that WGD is a single catastrophic event.

Timing patterns of copy number gains

To systematically explore the timing of copy number gains pan-cancer, we applied

mutational timing analysis to all 2,778 samples from 2,658 distinct donors across the PCAWG dataset (see **Supplementary Methods**). We find that chromosomal gains are typically acquired during the second half of clonal evolution (median value 0.76, IQR = 0.43-0.94), with systematic differences between tumour types (**Fig. 2a, Supplementary Fig. 1**). In glioblastoma, medulloblastoma and pancreatic neuroendocrine cancers, a substantial fraction of gains occurs early in mutational time. Conversely, in squamous cell lung cancers and melanomas, gains arise towards the end of the mutational time scale. Most tumour types, including breast, ovarian and colorectal cancer, show relatively broad periods of chromosomal instability, rather than staggered events throughout clonal evolution.

There are, however, certain tumour types with consistently early gains of specific chromosomal regions. Most pronounced is glioblastoma, where single copy gains of chromosomes 7, 19 and/or 20 are present in 90% of tumours (**Fig. 2a-b**). Strikingly, these gains are consistently timed within the first 10% of clonal mutational time. Similarly, the duplications leading to isochromome 17q in medulloblastoma are timed exceptionally early. Although less pronounced, gains of chromosome 18 in B-cell non-Hodgkin lymphoma, as well as gains of the q arm of chromosome 5 in clear cell renal cell carcinoma, often have a distinctively early timing within the first 50% of mutational time.

We observed that co-occurring gains in the same tumour often appear to occur at a similar time, pointing towards punctuated bursts of copy number gains involving the majority of gained segments (**Fig. 2c**). While this is expected in tumours with WGD (**Fig. 1b**), it may seem surprising to observe synchronous gains (defined as more than 80% of gained segments in a single event) in near-diploid tumours. Still, synchronous gains are frequent, occurring in a striking 58% (469/814) of informative near-diploid tumours, 61% more frequently than expected by chance ($p < 0.01$, permutation test; **Fig. 2d**). These data indicate that tumour evolution is often driven in short bursts involving multiple chromosomes, confirming earlier observations in breast cancer[16].

Timing of mutations in driver genes

As outlined above, point mutations can be qualitatively assigned to different time

categories, allowing the timing of driver mutations (**Fig. 1a, 3a**). Using a panel of 453 cancer driver genes[17], we find that the timing distribution of pathogenic mutations in the 50 most common drivers is predominantly clonal, and often early clonal (**Fig. 3a-b**). For example, *TP53* and *KRAS* are 5-9x more likely to be mutated in the early than in the late clonal stage. For *TP53*, this trend is independent of tumour type (**Fig. 3c**). Mutations in *PIK3CA* are 4x more frequently clonal than subclonal, while non-coding changes near the *TERT* gene are 8x more frequently early clonal than expected. In contrast, *SETD2* mutations are frequently subclonal, in agreement with previous reports[5]. Mutations in the non-coding RNA *RMRP* appear to be frequently late and subclonal.

Overall, common driver mutations predominantly occur early during tumour evolution. To understand how the entire landscape of all 453 driver genes changes over time, we calculated how the number of driver mutations relates to the number of driver genes in each of the evolutionary stages. This reveals an increasing diversity of driver genes mutated at later stages of tumour development: 50% of all early clonal driver mutations are found in only 12 different genes, whereas the corresponding proportion of late and subclonal mutations occur in approximately 39 and 36 different genes, respectively, a more than 3-fold increase (**Fig. 3d**). These results are consistent with previous findings in non-small-cell lung cancers[18], and suggests that, across cancer types, the very early carcinogenic events occur in a constrained set of common drivers, while a more diverse array of drivers is involved in late tumour development.

Relative timing of somatic driver events

Next, we sought to better understand the sequence and timing of events during tumour evolution by integrating the timing of driver point mutations and recurrent copy number changes across cancer samples. We calculated an overall probabilistic ranking of lesions, detailing whether each lesion occurs preferentially early or late during tumour evolution, by aggregating order relations between pairs of lesions from individual samples within each cancer type (**Supplementary Methods,** section 3.2, **Supplementary Fig. 2**).

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. 3e**). Whole-genome duplications have an intermediate ranking, indicating a variable timing, while many chromosomal gains and losses are typically late. These results are in agreement with the classical progression of *APC-KRAS-TP53* proposed by Vogelstein and Fearon[3], but add considerable detail.

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

Across cancer types, we typically find *TP53* mutations early, as well as losses of chromosome 17 (**Supplementary Fig. 1**). WGD events usually have an intermediate ranking and the majority of copy number changes occur after WGD. We also find that losses typically precede gains, and consistent with the results above, we find that common drivers typically occur earlier than rare drivers.

Timing of mutational signatures

Mutagenic processes acting on the tumour genome often leave characteristic signatures of their activity[19,20]. In order to quantify how these processes change over time, we estimated the intensity of active signatures within each sample, across the qualitative epochs of tumour evolution (early clonal, late clonal and subclonal). The changes in proportion of mutations associated with a given signature in each of these epochs provide a measure of the dynamics of relative signature activity (**Fig. 4**, **Supplementary Fig. 3**).

Overall, we find that signature activities typically change during clonal evolution by less than 30% (median fold change 0.98, IQR [0.70-1.36]), indicating that mutational

processes act at a rather constant rate during tumour progression. This is in contrast with the variation of signatures across patients, which varies 10 to 100-fold. There are, however, particular signatures that show consistent trends over time, both pan-cancer and within certain tumour types (**Fig. 4**). For example, the relative activity of the mutational signature associated with DNA damage caused by tobacco smoking (signature 4) decreases at least 1.2-fold in 70% of cancers where it is active clonally, consistent with previous reports in lung adenocarcinoma[21,22].

Other signatures, including UV light (signature 7) in melanoma (40% of samples with clonally active signature), and signature 12, of unknown aetiology, in liver cancer (83% of samples) show a similar ≥1.2-fold decrease in activity towards the later stages of clonal evolution (**Fig. 4**). We also observe that some signatures increase in late clonal evolution, most notably signatures 2 and 13, which are associated with the activity of APOBEC enzymes and increase by more than 1.2-fold in 58% of samples that have this signature. Similarly, the signature associated with *BRCA* mutations and defective double strand break repair (signature 3) increases in late clonal evolution in 35% of the samples where it is active. Similar trends also hold between clonal and subclonal phases of tumour evolution (**Supplementary Fig. 3**).

Chronological time estimates of whole genome duplications and subclonal diversification

Any changes in the mutation rate of cancers influence timing estimates made from mutational data. Due to increased proliferation and in some cases acquired hypermutation, one would generally expect an increase in the mutation rate (per year) in cancer, yet some mutational processes appear more variable than others.

The above analysis of signature changes revealed that the relative contribution of signature 1 usually decreases as other mutational processes become more active (**Fig. 4**). Mutational signature 1, characterised by CpG>TpG mutations, is a promising candidate for a clock-like process, as it is ubiquitously active in all tissues and has been described as correlating with age in normal tissues[23,24] and multiple tumour types[25]. The latter implies not only that it is fairly constant in a given cell lineage, but also that it varies little across patients. For the purpose of timing mutations in

chronological time, only the former property is required, as the age at diagnosis provides a reference by which relative timing estimates are scaled.

The acceleration of overall mutation rate and CpG>TpG rate can be directly estimated from sequencing data of matched primary and relapse samples from the same donor by comparing the rates of mutations that have accumulated between fertilisation and primary diagnosis to those accumulated between diagnosis and relapse. Suitable samples are publicly available for ovarian cancer[26], breast cancer[27] and acute myeloid leukaemia[28]. While for all point mutations, the median acceleration ranges between 3.3 for AML and 11.7 for ovarian cancer, CpG>TpG mutations display lower values and less variability (ranging from 2.8 to 6.7; **Fig. 5a**). To some extent this acceleration may be driven by treatment, but we may use it as a conservative reference for other tumour types.

Accounting for the acceleration above, we inferred the chronological time of whole-genome duplications based on CpG>TpG mutations (**Supplementary Methods**, section 5; **Fig. 5b**). While the typical timing of WGD is about one decade before diagnosis (assuming a 5x CpG>TpG mutation acceleration), we observe substantial variability among samples of a given tumour type, with many cases dating back more than two decades. Ovarian adenocarcinoma shows very early occurrences of WGD with approximately half of the samples having WGD more than two decades before diagnosis (**Fig. 5b**). A similar phenomenon is seen for breast adenocarcinoma. Without any acceleration, the estimated median occurrence of WGD would be 15-25yrs for the majority of cancer types; this value decreases with greater values of CpG>TpG acceleration (**Fig. 5c**).

We used a similar approach to calculate the timing of the emergence of the MRCA, and therefore the onset of subclonal diversification. The typical timing is considerably closer to diagnosis although, interestingly, there are also cases dating back more than ten years before diagnosis (**Fig. 5d**). We note, however, that timing the occurrence of the MRCA is more difficult, as it is not always possible to calculate the phylogenetic relationship between subclones. The MRCA may date back longer if subclones arise sequentially.

While the exact timing of individual samples remains challenging due to low

mutation numbers and unknown mutation rates for individual tumours, on average, a picture emerges where across tumour types, the median MRCA ranges between six months and six years before diagnosis, while WGD typically occurs 2-11 years before diagnosis (**Fig. 5e**). These findings dovetail with epidemiological observations: cancer generally arises past the age of 50[29], and the typical latency between carcinogen exposure and cancer detection, most notable in tobacco-associated cancers, is several years to multiple decades[30]. Furthermore the progression of most known precancerous lesions to carcinomas occurs usually over multiple years, if not decades[31-38]. The data presented here corroborate that these time scales hold also in cases without detectable premalignant conditions, raising hopes that these tumours could also be detected in precancerous stages.

Discussion

Taken together, these analyses begin to build an overall picture of tumour development. Across cancer types, early tumour development is characterised by mutations in a handful of canonical driver genes, and biallelic inactivation of tumour suppressor genes, such as *TP53*. Copy number gains during this time are relatively infrequent in many tumour types, but can be distinctive in others. Throughout the later stages of tumour evolution, increased genetic instability, a greater diversity of drivers, and an acceleration of mutational processes shape the final subclonal diversification.

Our combined approaches allow us to draw timelines of tumour development over different cancer types (**Fig. 6**; **Supplementary Fig. 1**). We see that many years before a tumour is diagnosed, endogenous and exogenous mutational processes have resulted in key driver mutations and chromosomal instability. An intriguing finding is that large somatic events, such as WGD, can occur decades before the appearance and diagnosis of a tumour. Thus, the process of tumour development may span an entire lifetime.

Our findings raise the possibility of early detection, if cells carrying early mutations can be detected and distinguished from cells not progressing further. The discovery of distinctive, early mutations in certain tumour types, such as gains of chromosomes 7,

losses of chromosome 10, and EGFR mutations in glioblastoma, and isochromosome 17q in medulloblastoma, begin to unveil possible candidate lesions.

Individual tumour types show characteristic sets of evolutionary trajectories, reflecting differences in the underlying biology of tumorigenesis (**Fig. 6**; **Supplementary Fig. 1**). Where applicable, these trajectories agree with previous studies of genomic aberrations acquired at different stages of tumour progression (e.g. in colorectal cancer[3]). Unlike most other cancers, high grade serous ovarian adenocarcinomas typically acquire chromosomal gains within the first half of clonal evolution (**Fig. 6d**). Our findings are consistent with these tumours being the most genomically unstable of all solid cancers[39], and with their high frequency of *TP53* and homologous recombination repair defects[40]. Both across and within cancer types, these typical evolutionary trajectories and their correlations with clinical features may provide an opportunity to develop prognostic markers and more effective therapies.

Our findings provide insight into the process of selection acting on tumours throughout their development. The genetic canalization in early tumour development, and increased diversity of driver mutations later in tumour evolution, is striking. It suggests a strong epistasis of fitness effects constraining evolution initially to a small set of mutational events that are able to initiate neoplastic transformation. Over time, as tumours evolve, the small- and large-scale somatic changes they subsequently accumulate propel them towards increasingly specialised developmental paths driven by individually rare, atypical driver mutations.

In summary, we present the first pan-cancer analysis of the evolutionary history of tumours. The timelines we derive from this analysis show that in a wide range of cancer types, tumour evolution often follows a typical pattern. This can begin decades before diagnosis, thus providing a window for early diagnosis and clinical intervention.

References

1 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23--28, doi:10.1126/science.959840 (1976).

2 Knudson, A. G. J. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820--823 (1971).

3 Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759--767, doi:10.1016/0092-8674(90)90186-I (1990).

4 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

5 Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

6 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).

7 Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* **21**, 751-759, doi:10.1038/nm.3886 (2015).

8 Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov* **5**, 1164-1177, doi:10.1158/2159-8290.CD-15-0369 (2015).

9 Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov* **1**, 137-143, doi:10.1158/2159-8290.CD-11-0028 (2011).

10 Purdom, E. *et al.* Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29**, 3113-3120, doi:10.1093/bioinformatics/btt546 (2013).

11 Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616-3627, doi:10.1182/blood-2013-08-518886 (2013).

12 Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression

and relapse. *Nature* **526**, 525-530, doi:10.1038/nature15395 (2015).

13 Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *bioRxiv* (2017).

14 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

15 McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

16 Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48**, 1119-1130, doi:10.1038/ng.3641 (2016).

17 PCAWG working group 2-5-9-14 (Analysis of mutations). *Manuscript in preparation* (2017).

18 Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).

19 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

20 PCAWG working group 7 (Mutation signatures and processes). *Manuscript in preparation* (2017).

21 McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* **7**, 283ra254, doi:10.1126/scitranslmed.aaa1408 (2015).

22 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, doi:10.1186/s13059-016-0893-4 (2016).

23 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

24 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

25 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet*, doi:10.1038/ng.3441 (2015).

26 Patch, A.-M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-494, doi:10.1038/nature14410 (2015).

27 Yates, L. R. & others. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **in press** (2017).

28 Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, doi:10.1038/nature10738 (2012).

29 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer Statistics, 2017. *CA Cancer J Clin* **67**, 7-30, doi:10.3322/caac.21387 (2017).

30 Thun, M. J., Henley, S. J. & Calle, E. E. Tobacco use and cancer: an epidemiologic perspective for geneticists. *Oncogene* **21**, 7307-7325, doi:10.1038/sj.onc.1205807 (2002).

31 Bostwick, D. G. & Qian, J. High-grade prostatic intraepithelial neoplasia. *Mod Pathol* **17**, 360-379, doi:10.1038/modpathol.3800053 (2004).

32 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, doi:10.1136/gut.2007.122739 (2007).

33 Gazdar, A. F. & Brambilla, E. Preneoplasia of lung cancer. *Cancer Biomark* **9**, 385-396, doi:10.3233/CBM-2011-0166 (2010).

34 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, doi:10.1002/cncr.21069 (2005).

35 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).

36 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, doi:10.1016/j.gtc.2015.02.005 (2015).

37 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478, doi:10.1038/nm.3733 (2014).

38 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

39 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

40 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

Figure Legends

Figure 1. Principles of timing mutations

(**a**) Principles of timing mutations based on deep whole genome sequencing. According to the clonal evolution model of cancer, tumour cells evolve in multiple selective sweeps. During some of these sweeps, copy number gains are acquired, which can be used for timing analyses (green and purple epochs). Mutations acquired after the last clonal expansion are present in distinct subclonal populations (red epoch). The number of sequencing reads reporting point mutations can be used to discriminate variants as early or late clonal (green/purple) in cases of specific copy number gains, as well as clonal (blue) or subclonal (red) in cases without (right). The distribution of the number of early and late clonal mutations carries information about the timing of the copy number gains with the exact relation depending on the resulting copy number configuration (bottom). (**b**) Example case illustrating the annotation of point mutations based on the variant allele frequency (VAF, top) and copy number configuration (middle), each shown as a function of genomic coordinate (x-axis). The resulting timing estimates for each copy number segment are shown at the bottom, indicating that all segments were gained at a similar time (whole genome duplication).

Figure 2. Pan-cancer timing patterns of arm-level gains

(**a**) Overview of timing arm-level copy number gains across different cancer types. Depicted are the smoothened histograms (y-axes; scale bar 5% recurrence) of the timing estimates of large gains at decile resolution (x-axes), split by tumour type and chromosome on which gains are detected. (**b**) Heatmaps representing timing estimates of gains on different chromosome arms (x-axis) for individual samples (y-axis) for selected tumour types. (**c**) Two near-diploid example cases illustrating synchronous gains with a single peak in amplification activity (top) and asynchronous gains with multiple amplification periods (bottom). (**d**) Distribution of synchronous and asynchronous gain patterns across samples, split by whole genome duplication status (left). Uninformative samples carry too few or too small gains to be timed accurately. Systematic permutation tests reveal a 61% enrichment of synchronous gains in near-diploid samples (right).

Figure 3. Timing of driver mutations and relative ordering of somatic events

(**a-d**) **Timing of driver point mutations.** (**a**) Top: distribution of point mutations over different mutations periods in 2,583 samples from unique donors. Middle: timing distribution of driver point mutations in the 50 most recurrent lesions. 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 (green, purple) or clonal versus subclonal (blue, red). Odds ratios overlapping 1 in less than 5% of bootstrap samples are considered significant and have been coloured. (**c**) Relative timing of *TP53* mutations across cancer types, coloured as in (**b**). (**d**) Estimated number of unique lesions (genes) contributing 50% of all driver mutations in different timing epochs. Error bars denote the range between 0 and 1 pseudocounts. (**e**, **f**, **g**) **Relative ordering of somatic events.** Preferential ordering diagrams of somatic copy number events and driver point mutations within tumour types, for (**e**) colorectal adenocarcinoma, (**f**) pancreatic neuro-endocrine cancer and (**g**) glioblastoma. Probability distributions show the uncertainty of timing for specific events in the cohort. Events with odds above 10 (either earlier or later) are highlighted.

Figure 4. Timing of signatures

(**a**) Fold changes in signature exposures between early and late clonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (**b**) Fold changes in signature exposures in individual tumours (early *vs*. late clonal). Within cancer types, tumours are ordered according to hierarchical clustering. White indicates inactive signatures.

Figure 5. Real-time estimation of mutational landmarks

(**a**) Mutation rate acceleration inferred from paired samples. CpG>TpG mutations (right) display a lower acceleration rate compared to all point mutations (left). (**b**)

Time of occurrence of whole genome duplications in individual patients, split by tumour type, based on CpG>TpG mutations and patient age. Results are shown for a 5x acceleration of the mutation rate. (**c**) Median time of WGD occurrence per cancer type, as a function of CpG>TpG acceleration. (**d**) Timing of subclonal diversification using CpG>TpG mutations in individual patients. (**e**) Comparison of inferred median occurrence of WGD and subclonal diversification.

Figure 6. Cancer timelines

Typical timelines of tumour development, for (**a**) glioblastoma, (**b**) colorectal adenocarcinoma, (**c**) squamous cell lung cancer, (**d**) ovarian adenocarcinoma, and (**e**) pancreatic adenocarcinoma. Each timeline represents the length of time, in years, between the fertilised egg and the median age of diagnosis per cancer type. Point estimates for major events, such as WGD and the emergence of the MRCA are used to define early, intermediate, late and subclonal stages of tumour evolution approximately in chronological time. Driver mutations and copy number aberrations are shown in each stage according to their preferential timing, as defined by relative ordering. Mutational signatures that fluctuate during tumour evolution, either considerably (median change +/- 20%), or consistently (75% samples change in the same direction) are annotated as well.

Methods

Timing of gains

We used three related approaches to calculate the timing of copy number gains (see **Supplementary Methods**, section 1). In brief, the common feature is that the expected variant allele frequency of a mutation is related to the underlying number of alleles carrying a mutation according to the formula

$$E[X] = n\, m\, f \,/\, [N\,(1-\rho) + C\,\rho]$$

Here 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 the total copy number of the tumour and ρ the purity of the sample.

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

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

Copy number 2+2: $T = 2\, n_2 \,/\, (2n_2 + n_1)$

Copy number 2+0: $T = 2\, n_2 \,/\, (2n_2 + n_1)$

Here 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.

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 Methods**, section 1.2. Mutations with $f = 1$ are denotes as clonal, and mutations with $f < 1$ as *subclonal*. Mutations with $f = 1$ and $m > 1$ are denoted as *early clonal* (coamplified). 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*] (**Supplementary Methods**, section 1.2.)

Timing of driver mutations

A catalogue of driver point mutations was provided by the PCAWG Drivers and Functional Interpretation Group[17]. The timing category was calculated as above. From the four timing categories, 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 in the samples with each particular driver. The background distribution of these odds ratios was assessed with 1000 bootstraps (**Supplementary Methods**, section 3.1.)

Integrative timing

For each pairs of driver point mutations and recurrent copy number variants it was established what the ordering of the given pair was (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 Methods**, section 3.2.

Timing of mutational signatures

Mutational trinucleotide substitution signatures, as defined by the PCAWG Mutational Signatures Working Group[20], were refit to samples with observed

signature activity, after splitting point mutations into either of the 4 timing categories. Time-resolved exposures were calculated using non-negative linear least squares. Full details are given in **Supplementary Methods**, section 4.

Real-time estimation of copy number gains

For tumours with multiple 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.

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 0.66 to 1.0 of relative time and averaged the results. 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 Methods**, section 5.

Supplementary Figure Legends

Supplementary Figure 1. Summary of all results obtained per cancer type

(**a**) Clustered heatmaps 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 per cancer type. (**c**) Distribution of mutations across early clonal, late clonal and subclonal stages, for the most common driver genes per cancer type. 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 *vs*. subclonal stages. Blue indicates a signature decrease and red an increase in subclonal from clonal mutations. (**f**) Typical timeline of tumour development, per cancer type.

Supplementary Figure 2. Correlation between league model and Bradley-Terry model order of events.

The two approaches for determining the order of recurrent somatic mutations and copy number events are compared directly for each tumour type. We show how the order derived from the league model compares to that derived from the Bradley-Terry model, quantified by Spearman's rank correlation coefficient.

Supplementary Figure 3. Timing of signatures

(**a**) Fold changes in signature exposures between clonal and subclonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (**b**) Fold changes in signature exposures in individual tumours (clonal *vs*. subclonal). Within cancer types, tumours are ordered according to hierarchical clustering. White indicates inactive signatures.

Author contributions

MG, CJ, IL, SG, PA, DR, DGL, PTS and PVL performed timing of point mutations and copy number gains. SG and MG performed qualitative timing of driver point mutations. IL, TJM, DR, DGL, DCW and GG performed relative timing of somatic driver events and implemented integrative models. CJ, YR, PVL and QDM performed timing of mutational signatures. MG performed real-time estimation of whole-genome duplication and subclonal diversification. CJ, MG, IL, YR, DR and PVL constructed cancer timelines. MG, CJ, IL, SCD, SG, TJM, YR, PA, JD, PCB, DDB, VM, QDM, PTS, DCW and PVL interpreted the results. SCD, IL, JW, AD, IVG, KeY, GM, MP, SM, ND, KaY, SSe, KH, MT, JD, DGL, DR, JL, MC, SCS, YJ, FM, VM, HZ, WW, QDM, DCW and PVL performed subclonal architecture analysis. SCD, IL, KK, VM, MP, XY, DGL, SSc, RB, MI, MS, DCW and PVL performed copy number analysis. JW, SCD, IL, KH, DGL, KK, DR, DCW, QDM and PVL derived a consensus of copy number analysis results. KaY, MT, AD, SCD, IL, DCW, MG, PVL, QDM and WW derived a consensus of subclonal architecture results. YF and WW contributed to subclonal mutation calls. PTS, DCW and PVL coordinated the study. MG, CJ, PTS, YR, IL, QDM, DCW and PVL wrote the manuscript.

Acknowledgements

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). MT and JD are postdoctoral fellows supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 747852-SIOMICS and 703594-DECODE). JD is a postdoctoral fellow of the FWO. FM, GM and KeY would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. GM, KeY and FM were funded by CRUK core grants C14303/A17197 and A19274. SSe and YJ are supported by NIH R01 CA132897. HZ is supported by grant NIMH086633 and an endowed Bao-Shan Jing Professorship in Diagnostic Imaging. WW is supported by the U.S. National

Cancer Institute (1R01 CA183793 and P30 CA016672). PTS was supported by U24CA210957 and 1U24CA143799. DCW is funded by the Li Ka Shing foundation. PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute.

Members of the PCAWG Evolution and Heterogeneity Working Group

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

[1]The Francis Crick Institute, London NW1 1AT, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; [3]Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom; [6]University of Toronto, Toronto, ON M5S 3E1, Canada; [7]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97231, USA; [8]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; [9]Peter MacCallum Cancer Centre, Melbourne, VIC 3052, Australia; [10]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [11]University of Melbourne, Melbourne, VIC 3010, Australia; [12]University of Cologne, 50931 Cologne, Germany; [13]Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium; [14]Simon Fraser University, Burnaby, BC V5A1S6, Canada; [15]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; [16]Heidelberg University, 69120 Heidelberg, Germany; [17]The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [18]Weill Cornell Medicine, New York, NY 10065, USA; [19]New York Genome Center, New York, NY 10013, USA; [20]NorthShore University HealthSystem, Evanston, IL 60201, USA; [21]The

University of Chicago, Chicago, IL 60637, USA; [22]University of California Santa Cruz, Santa Cruz, CA 95064, USA; [23]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, United Kingdom; [24]University of Cambridge, Cambridge CB2 0QQ, United Kingdom; [25]University of Helsinki, 00014 Helsinki, Finland; [26]Carleton College, Northfield, MN 55057, USA; [27]Princeton University, Princeton, NJ 08540, USA; [28]Indiana University, Bloomington, IN 47405, USA; [29]Baylor College of Medicine, Houston, TX 77030, USA; [30]University of Glasgow, Glasgow G12 8RZ, United Kingdom.

*: These authors contributed equally

[#]: These authors jointly directed the work

Figure 1. Principles of timing mutations

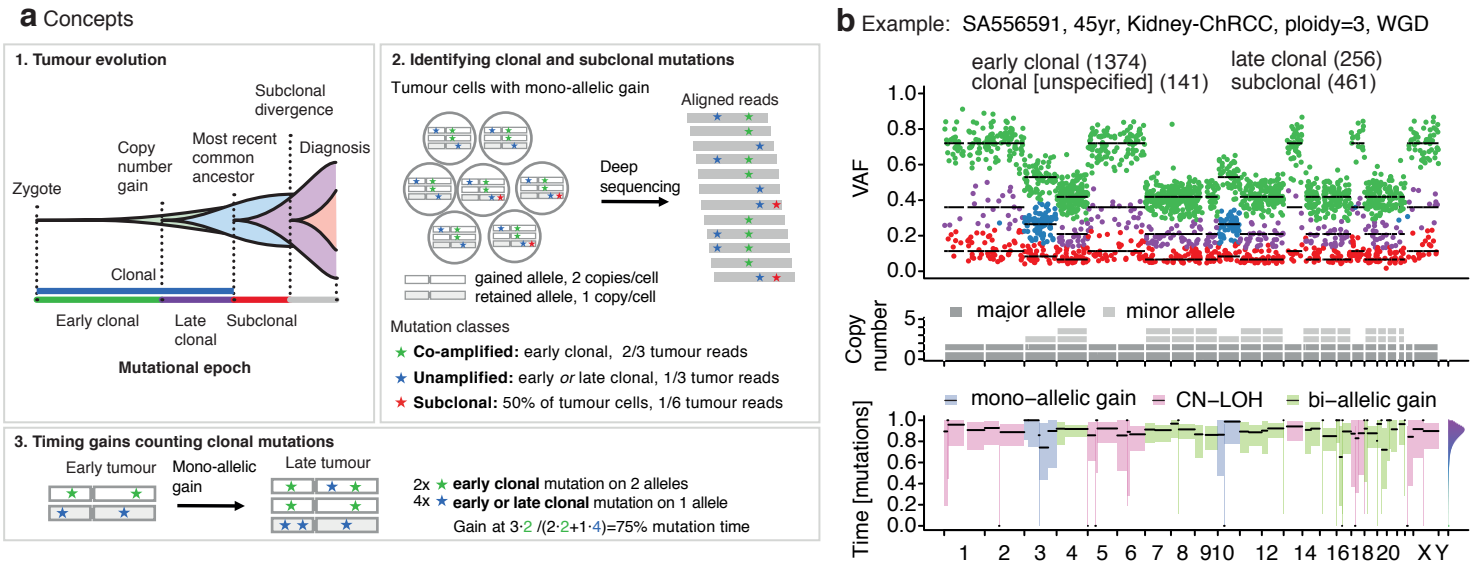

Figure 2: Pan-cancer timing patterns of arm-level gains

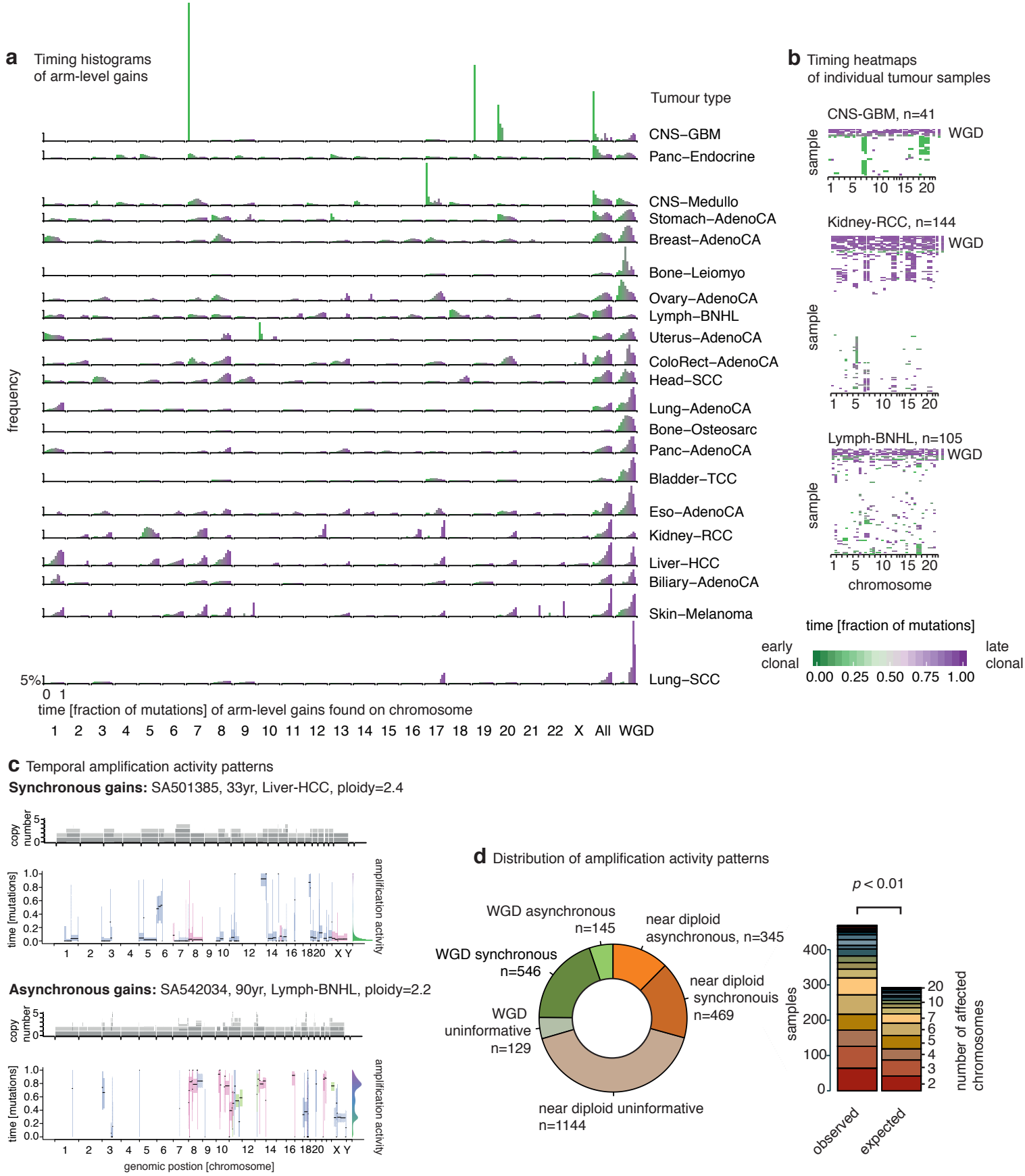

Figure 3: Timing of driver mutations and relative ordering of somatic events

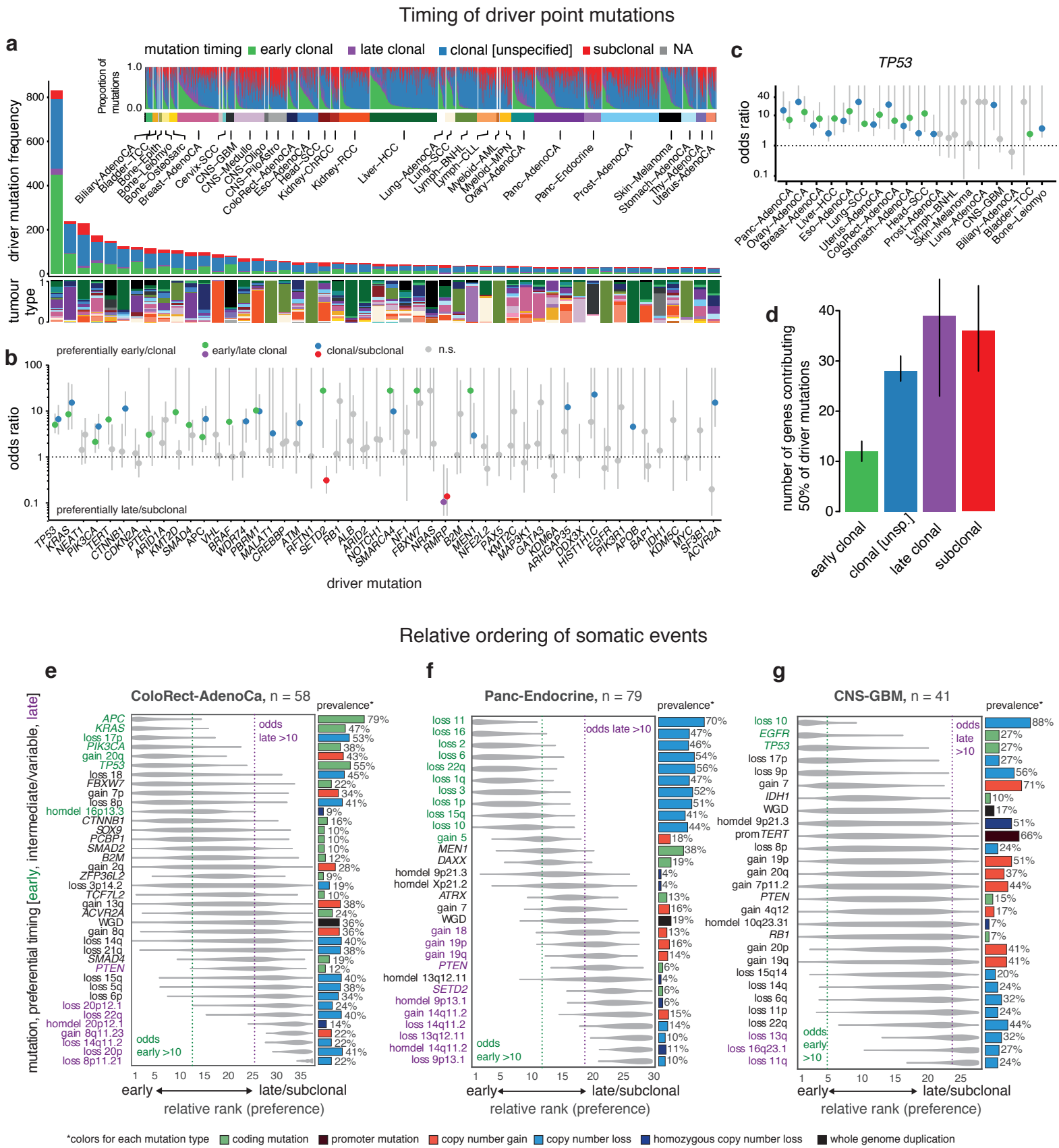

Figure 4: Evolution of signatures: early clonal vs late clonal

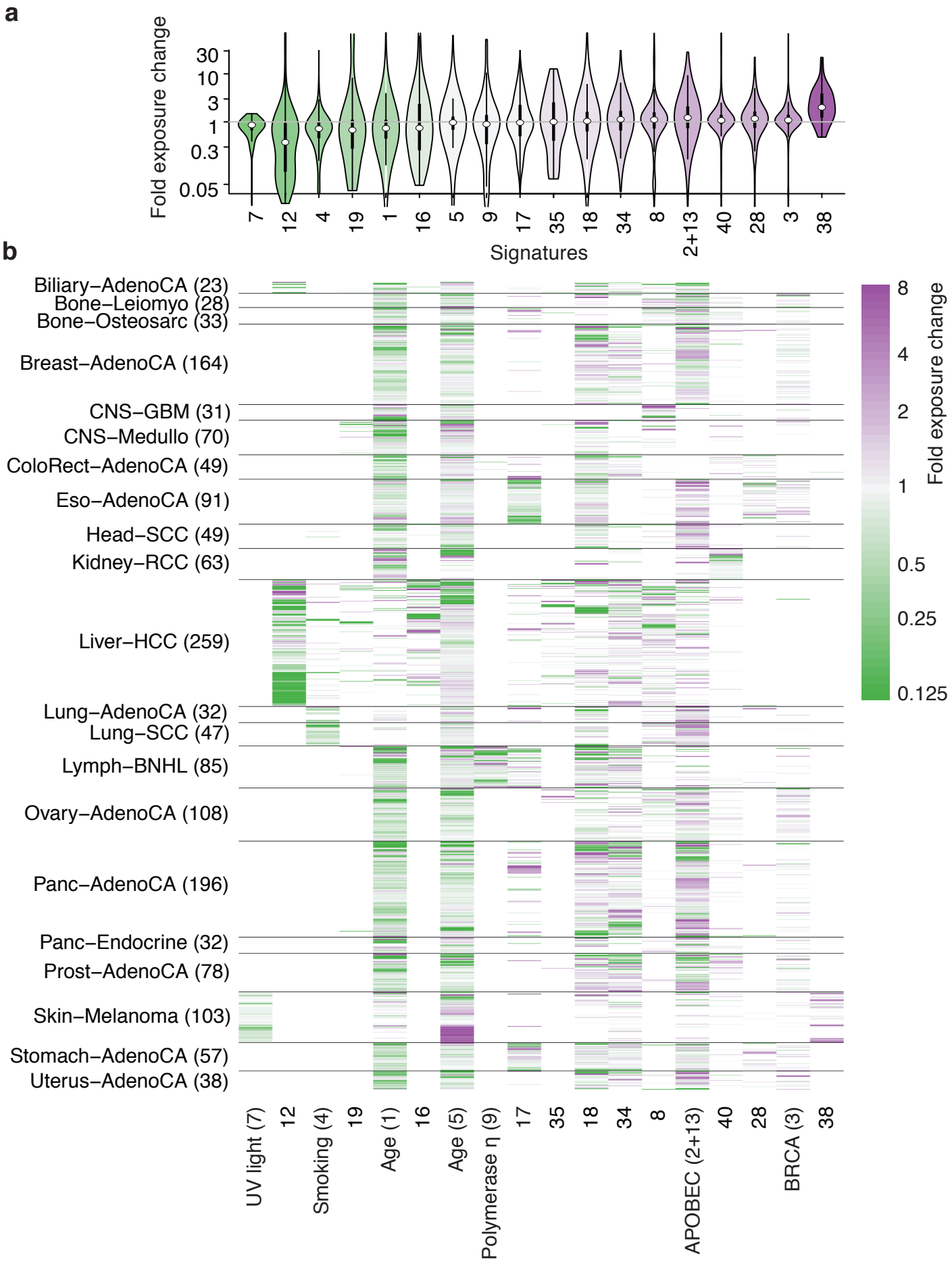

Figure 5: Real-time estimates

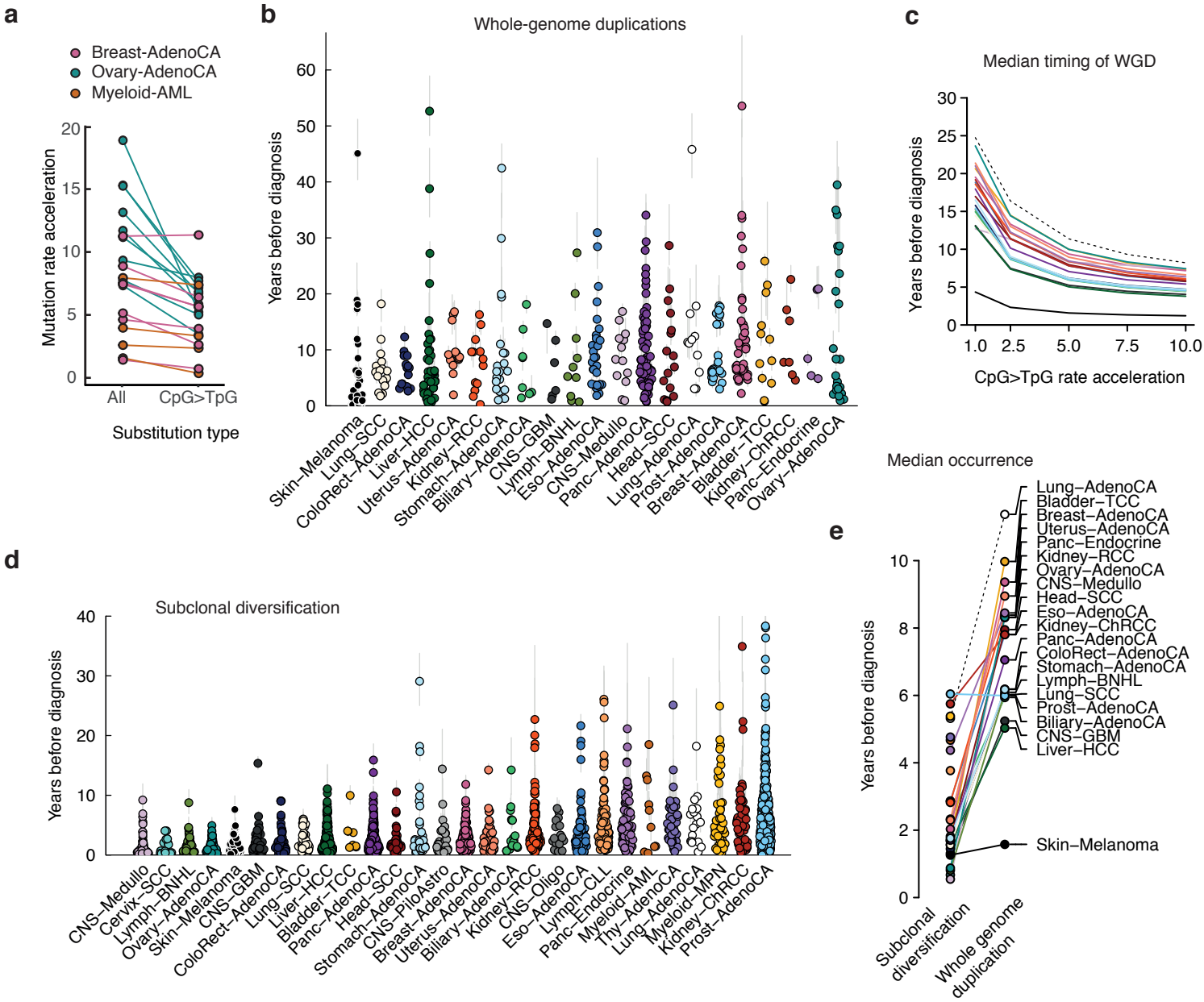
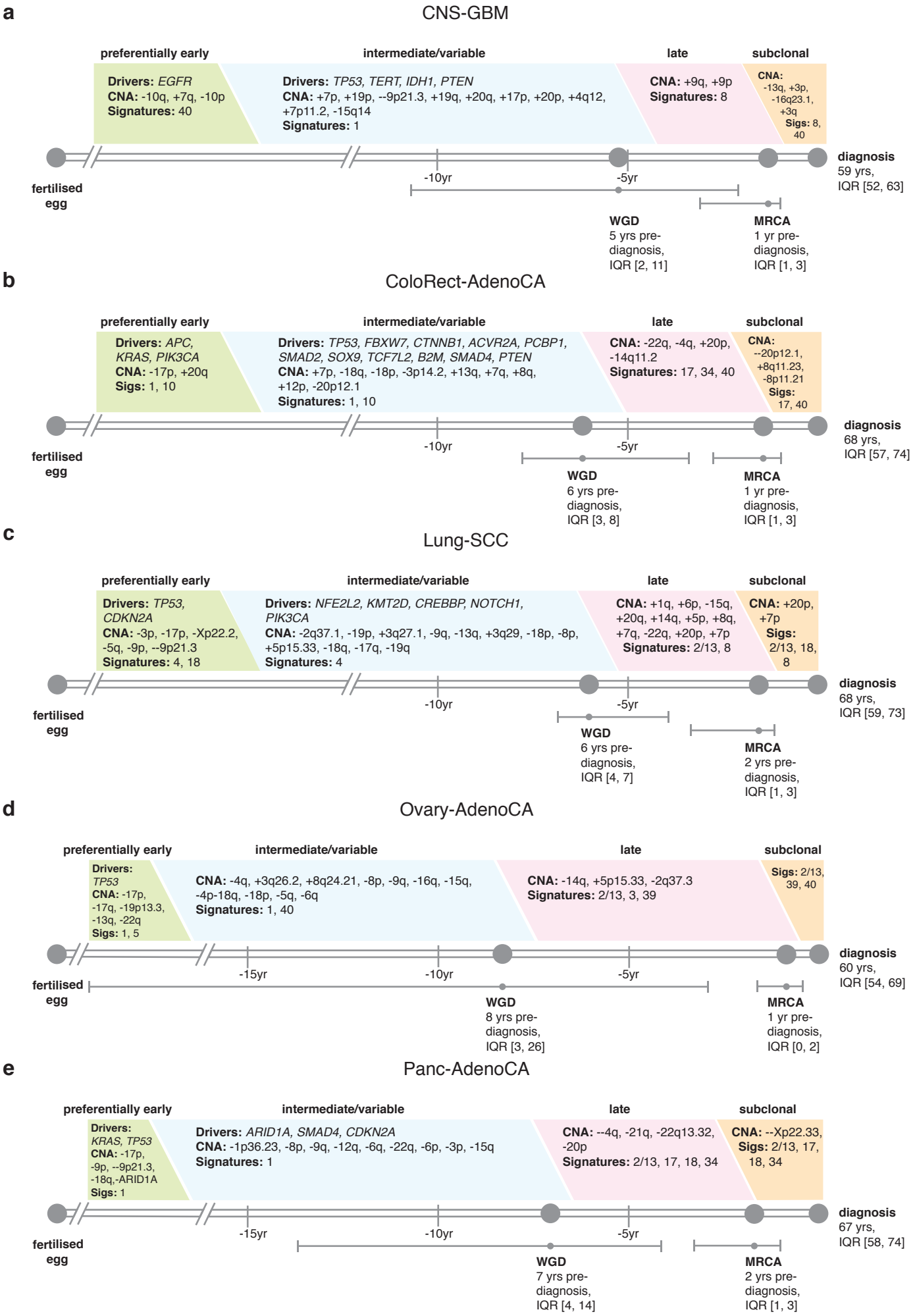
Figure 6: Oncogenic timelines