nature genetics

Article

https://doi.org/10.1038/s41588-024-01949-7

Mapping extrachromosomal DNA amplifications during cancer progression

Received: 13 December 2023

Accepted: 13 September 2024

Published online: 14 October 2024

Check for updates

Hoon Kim[1,2,3,15,16] ✉, Soyeon Kim[1,4,15], Taylor Wade[5], Eunchae Yeo[2], Anuja Lipsa[6], Anna Golebiewska[6], Kevin C. Johnson[4], Sepil An[1], Junyong Ko[7], Yoonjoo Nam[1], Hwa Yeon Lee[8], Seunghyun Kang[1], Heesuk Chung[1], Simone P. Niclou[6,9], Hyo-Eun Moon[10], Sun Ha Paek[10,11], Vineet Bafna[12,13], Jens Luebeck[12] & Roel G. W. Verhaak[4,14,16] ✉

To understand the role of extrachromosomal DNA (ecDNA) amplifications in cancer progression, we detected and classified focal amplifications in 8,060 newly diagnosed primary cancers, untreated metastases and heavily pretreated tumors. The ecDNAs were detected at significantly higher frequency in untreated metastatic and pretreated tumors compared to newly diagnosed cancers. Tumors from chemotherapy-pretreated patients showed significantly higher ecDNA frequency compared to untreated cancers. In particular, tubulin inhibition associated with ecDNA increases, suggesting a role for ecDNA in treatment response. In longitudinally matched tumor samples, ecDNAs were more likely to be retained compared to chromosomal amplifications. EcDNAs shared between time points, and ecDNAs in advanced cancers were more likely to harbor localized hypermutation events compared to private ecDNAs and ecDNAs in newly diagnosed tumors. Relatively high variant allele fractions of ecDNA localized hypermutations implicated early ecDNA mutagenesis. Our findings nominate ecDNAs to provide tumors with competitive advantages during cancer progression and metastasis.

Disease progression, including metastasis, is a leading cause of death from cancer as tumors acquire resistance and become increasingly less responsive to therapies[1,2]. Characterizing the genomic features of primary untreated and metastatic treated tumors is critical to improving our understanding of the processing driving cancer progression[3,4]. Cancer is driven by genomic alterations, including focal DNA amplifications, in which DNA segments containing oncogenes or oncogenic regulatory elements are multiplied, resulting in oncogene transcription and activation[5]. Amplifications may occur through mechanisms tethered to chromosomes, forming homogeneously staining regions (HSRs), or by excising and circularizing DNA segments to form extrachromosomal DNA (ecDNA) elements[6,7]. HSRs and ecDNAs both create gene amplification, but their functional consequences may vary[8,9]. EcDNAs replicate with the linear genome but lack centromeres, resulting in uneven segregation and enabling rapid accumulation of ecDNAs in tumor cell nuclei[9,10]. If the ecDNA endows the tumor cell with a competitive advantage, cells containing ecDNAs undergo selection, creating a dominant tumor cell clone driven by an ecDNA-activated oncogene[11]. The ecDNAs are detected in most human cancer types at the time of diagnosis and are enriched in poor prognosis tumor types such as glioblastoma, sarcoma and esophageal carcinoma[8]. However, the role of ecDNAs in advanced cancers remains unclear.

The genes carried on or activated by ecDNAs include *ERBB2*, *EGFR* and *CDK4*, which are targets of commonly used inhibitors for the treatment of patients with cancer. In addition, oncogenes that are considered undruggable are detected on ecDNAs, such as *MYC*, *TERT* and *MCL1*. In fact, all genes known to be focally amplified in cancer are detected on ecDNAs in some tumors[8,12,13]. The discovery of ecDNA clusters that appear to function as hubs where transcriptional

A full list of affiliations appears at the end of the paper. ✉e-mail: wisekh@skku.edu; roel.verhaak@yale.edu

https://doi.org/10.1038/s41588-024-01949-7

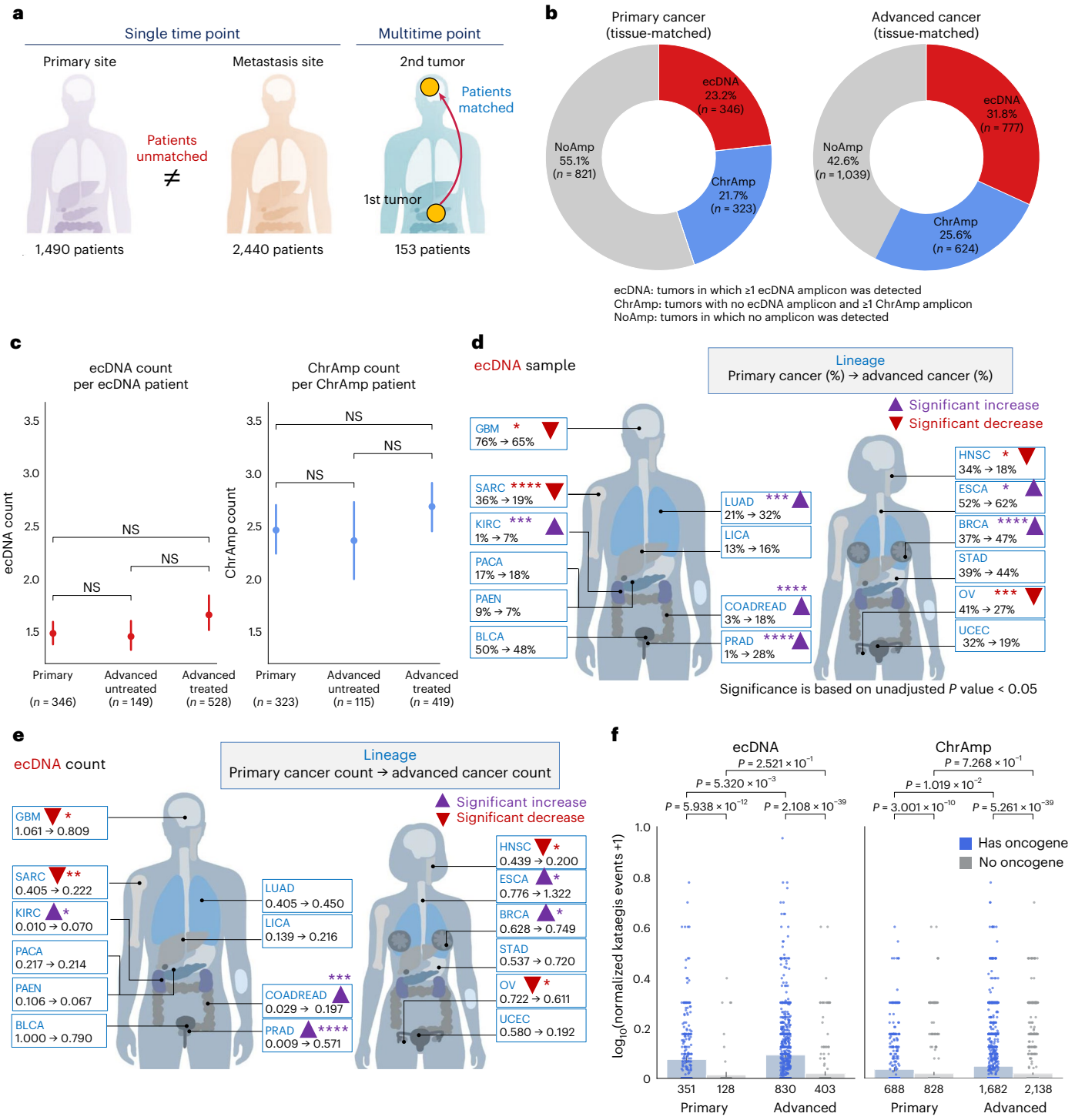


Fig. 1 | Sample classification. a, Schematic dataset overview. **b**, Overview of sample classification for 1,490 patients in the primary cancer cohort and 2,440 patients in the advanced cancer. Only tumor types with at least 20 patients in each cohort were included. **c**, Average number of ecDNA and ChrAmp amplicons detected per ecDNA patient and ChrAmp patient, respectively. Tumor lineages represented by at least 20 tumors in both cancer cohorts are included. Numbers in parentheses indicate the number of patients. Points represent mean values, and error bars show a 95% CI. *P* values were computed using a two-sided Mann–Whitney *U* test. **d**, Percentage of ecDNA samples. **e**, The average number of distinct ecDNA amplicons per sample in primary and advanced cancer cohorts, showing tumor lineage represented by at least 20 tumors in both cohorts. *P* values were computed using a one-sided binomial test with the ecDNA-carrying tumor fraction in the primary cancer cohort as a null probability in **d** and using a one-sided Mann–Whitney *U* test in **e** where not significant unless noted otherwise. **f**, Number of kataegis events normalized by the number of intervals present on ecDNA or ChrAmp amplicons in the primary and advanced cohorts, respectively. Numbers indicate the number of amplicons. Bars represent mean values, and error bars show 95% CIs. *P* values were computed using a two-sided Mann–Whitney *U* test. Asterisks indicate level of significance: $*1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2}$, $**1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2}$, $***1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3}$ and $****P \leq 1.00 \times 10^{-4}$. NS, not significant; GBM, glioblastoma multiforme; SARC, sarcoma; KIRC, kidney renal clear cell carcinoma; PACA, pancreatic cancer; PAEN, pancreatic cancer endocrine neoplasms; BLCA, bladder urothelial carcinoma; LUAD, lung adenocarcinoma; LICA, liver cancer; COADREAD, colorectal cancer; PRAD, prostate adenocarcinoma; HNSC, head and neck squamous cell carcinoma; ESCA, esophageal carcinoma; BRCA, breast invasive carcinoma; STAD, stomach adenocarcinoma; OV, ovarian serous cystadenocarcinoma; UCEC, uterine corpus endometrial carcinoma.

machinery is assembled and shared[9,14], the absence of centromeres that results in uneven segregation[11,15], the detection of ecDNA sequences in micronuclei[16,17] and the enrichment of enhancer elements on ecDNA molecules[18,19] contribute to the hypothesis that proteins regulating ecDNA-related processes may represent potent drug targets. Effective targeting of ecDNA elements requires understanding the role of ecDNA during cancer progression.

Here we have compared ecDNA frequencies and properties in cancers at the time of diagnosis and at later stages of disease to evaluate whether ecDNAs act as drivers of tumor evolution[11]. We determined the presence of ecDNAs through a computationally intensive and standardized analysis pipeline to uniformly process 8,060 whole-genome sequencing (WGS) datasets generated from biopsy specimens obtained from patients at cancer diagnosis and in patients with advanced pretreated and/or metastatic cancer, including 231 cases with multiple time-separated specimens.

ecDNAs are frequently detected in advanced tumors

We determined the incidence of ecDNA in progressed tumors through analysis of WGS datasets from 4,170 advanced cancer samples, derived from 4,170 patients, available through the Hartwig Medical Foundation (HMF)[20]. The HMF cohort included tumors from 2,333 pretreated patients, 1,191 untreated patients and 646 patients with unknown treatment status. We compared HMF results with those derived from analyzing the whole genomes of 3,464 newly diagnosed tumors and 226 pretreated tumors from The Cancer Genome Atlas–the International Cancer Genomics Consortium (TCGA–ICGC)[8] and 100 matching primary-recurrent pairs from the Glioma Longitudinal Analysis (GLASS) consortium[21]. The datasets were analyzed using AmpliconSuite-pipeline (v.0.1344.2) to detect focally amplified genomic loci and reconstruct the structures of the resulting amplicons from the whole-genome sequences from all 8,060 samples. The AmpliconSuite-pipeline includes the AmpliconArchitect[22] method to derive amplicon structures and the AmpliconClassifier to assign amplicons to an amplicon class (Supplementary Table 1)[23]. Amplicons carrying a circular amplicon structure signature were classified as ecDNA, and noncircular amplicons were grouped into the chromosomal amplification (ChrAmp) class[23]. In total, across 8,060 tumors, we detected 2,602 ecDNA amplicons and 8,594 ChrAmp amplicons. We further assigned sample-level classes, labeling tumors containing at least one ecDNA amplicon as ecDNA and samples with at least one noncircular amplicon as ChrAmp. Tumors lacking amplicons were labeled 'no focal somatic copy-number amplification' (NoAmp).

To be able to evaluate ecDNA frequencies between cohorts, we determined whether tumor purity and sequencing depth impacted the sensitivity of amplicon detection. We observed that a reduced number of ecDNAs were detected in samples with an average coverage of less than ten times (Extended Data Fig. 1a). Additionally, we found a significant difference in ecDNA frequency between ICGC and HMF samples in tumor purity bins 0.3–0.4 and 0.4–0.5 (Extended Data Fig. 1b). Comparisons in the TCGA cohort were limited by low sample numbers, following filtering of the <10× samples. Based on this observation, we additionally removed samples with tumor purity less than 0.4 from comparisons between cohorts. As a result, 2,196 TCGA–ICGC and 3,045 HMF tumors passed all filtering criteria. These samples were then used to construct a tissue-matched primary cancer cohort (n = 1,490) consisting of newly diagnosed and untreated TCGA–ICGC tumors and an advanced cancer cohort (n = 2,440) comprising metastatic and/or pretreated tumors from TCGA–ICGC and HMF, by including only tumor types represented by at least 20 samples in both primary and advanced cohorts (Fig. 1a and Extended Data Fig. 1c). After applying the same filters on 508 paired primary and recurrent/metastatic specimens, a longitudinal cohort consisting of 306 multitime point samples from 153 patients was created across TCGA, HMF and GLASS cohorts (Extended Data Fig. 1d).

At least one ecDNA was detected in 346 (23.2%) tumors from the primary cancer cohort and 777 tumors (31.8%) of the advanced cancer cohort (Fig. 1b and Extended Data Fig. 2a). A significantly larger fraction of the advanced cancer cohort harbored ecDNA and ChrAmp amplifications, and the average number of ecDNAs and ChrAmp amplicons per tumor in both amplicon classes was comparable between cohorts (Fig. 1c). We performed a resampling analysis in which tumor-type distribution was equal between cohorts, which confirmed that the increase in ecDNA and ChrAmp frequencies in advanced cohort tumors was independent of tumor lineage (Extended Data Fig. 2b). We confirmed high frequencies of samples containing ecDNA amplicons in glioblastomas (76%), esophageal carcinoma (52%) and bladder carcinoma (50%) cancers from the primary cancer cohort (Fig. 1d)[8]. The fraction of ecDNA samples and the average number of ecDNAs per sample significantly increased in the advanced cancer cohort clear cell renal and esophageal carcinoma, colorectal, prostate and breast cancer (Fig. 1e). In contrast, we observed a significant decrease in ecDNA sample fraction and ecDNA count in glioblastoma, sarcoma, head and neck and ovarian carcinoma. ChrAmp sample fraction and ChrAmp amplicon counts were observed to follow similar patterns (Extended Data Fig. 2c–e). These observations suggested that the driving roles of ecDNA and chromosomal amplicons may vary by tumor lineage.

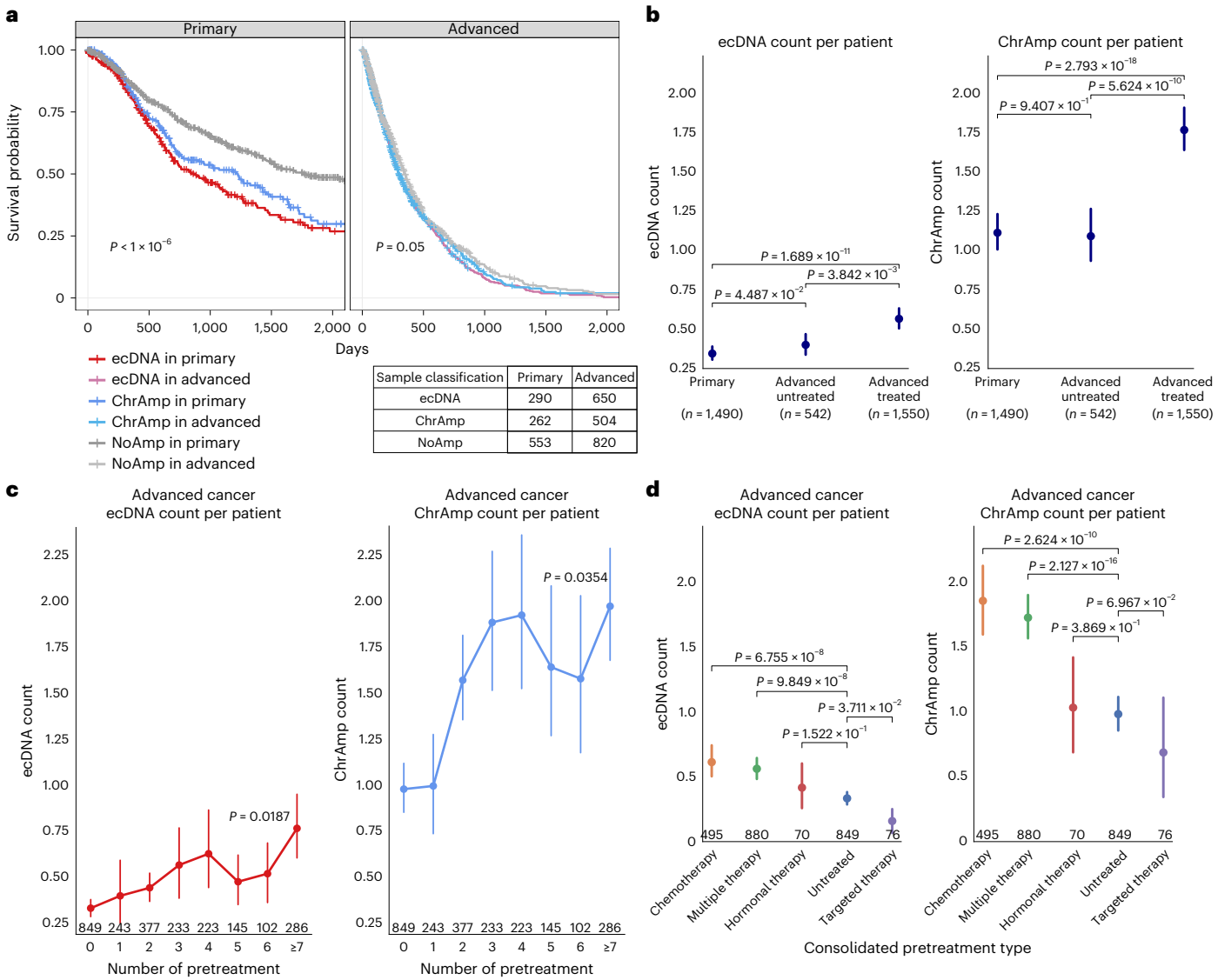
We evaluated the genomic characteristics of amplicons and found that the presence of an oncogene on the amplicon is the major determinant of amplicon complexity, which is a composite value based on the distribution of copy numbers assigned to reconstructions of the focal amplification's genome structure and the total number of genomic segments comprising an amplicon[23]. This was true for both ecDNA and ChrAmp (Extended Data Fig. 3a–c). Amplicon complexity, copy number and size did not significantly differ between primary and advanced cancer cohorts. Increased genome ploidy, whole-genome duplication and microsatellite instability but not homologous recombination associated with higher rates of ecDNA and contributed to the increased rates of ecDNA in the advanced cohort (Extended Data Fig. 3d–g and Extended Data Fig. 4a–d). The observed increased frequency of ecDNA in tumors of the advanced cohort is thus, in part, explained by the higher levels of ploidy and whole-genome duplication.

Localized hypermutation (kataegis) has been reported to occur frequently on ecDNAs in primary tumors[24,25]. We confirmed the frequent co-occurrence of kataegis on ecDNA and ChrAmp amplicons in primary cancer tumors (Fig. 1f). As localized hypermutations often happen in the context of single- and double-strand DNA break repair[26], we normalized the frequency of clustered mutation events by the number of amplicon intervals. Kataegic clustered mutation events were detected at significantly higher rates in oncogene-containing but not nononcogenic ecDNAs, from the advanced cancer cohort and relative to the primary cancer cohort (Extended Data Fig. 4e). The significant difference in kataegis frequency was also observed among breast cancers, the largest cohort of a single tumor type within our datasets (Extended Data Fig. 4f). Our results suggest that ecDNAs containing oncogenes and kataegis are most likely to be detected as tumors progress.

Clinical associations of ecDNA across cancers

We previously showed that the presence of an ecDNA amplicon is associated with poor prognosis in newly diagnosed tumors[8]. We confirmed this association in the primary and advanced cancer cohorts (Fig. 2a). A multivariate analysis that additionally considered primary tumor location, primary versus advanced cohort, sex, age across multiple bins, whole-genome doubling status, microsatellite instability status, homologous recombination status and tumor stage showed that the presence of ecDNA was associated with an increase hazard ratio (P < 0.001 ecDNA versus NoAmp, P = 0.002 ChrAmp versus NoAmp; P values by multivariate cox proportional-hazard model; Extended Data Fig. 5a).

https://doi.org/10.1038/s41588-024-01949-7



Sample classification	Primary	Advanced
ecDNA	290	650
ChrAmp	262	504
NoAmp	553	820

Fig. 2 | Clinical associations. a, Five-year Kaplan–Meier survival curves by amplification category using patients. The *P* value derived from comparing the survival curves was based on a log-rank test in the primary and advanced cohorts, separately. **b**, Distribution of the number of distinct ecDNA and ChrAmp amplicons by pretreatment status across primary, untreated advanced cancers and pretreated advanced cancer tumors. Pretreated advanced cancer tumors show a significantly higher number of distinct ecDNAs and ChrAmps per tumor compared to primary cancer or untreated advanced cancer tumors (two-sided Mann–Whitney *U* test). *Y* axis represents the number of distinct ecDNA and ChrAmp amplicons detected per tumor. Numbers indicate patient counts. All tumors with available pretreatment information were included in the analysis. Points represent mean values, and error bars show 95% CIs. **c**, Distribution of the number of distinct ecDNA and ChrAmp amplicons by the number of pretreatments received across pretreated HMF advanced cancers. *P* value was calculated using a two-sided Mann–Kendall trend test. Points represent mean values, and error bars show a 95% CI. Only patients with available clinical information were included. Numbers indicate the number of patients. **d**, Distribution of the number of distinct ecDNA and ChrAmp amplicons by different prebiopsy treatment types in the advanced cancer cohort. 'Untreated' category only includes tumors from the advanced cohort. Number of patients per category is shown on the bottom. Only treatment types used in more than 50 patients are shown. *P* values were calculated using a two-sided Mann–Whitney *U* test. Points represent mean values, and error bars show a 95% CI.

Many but not all patients included in HMF have previously undergone cancer therapy, which can alter the genomic properties of the tumor[27]. Untreated HMF patients ($n = 542$) were in majority newly diagnosed with metastatic cancer[4]. We observed that the ecDNA count per tumor was significantly higher in untreated HMF tumors compared to the primary cancer cohort (0.34, 95% confidence interval (CI): 0.30, 0.39 versus 0.4, 95% CI: 0.33, 0.47, $P = 0.045$, Mann–Whitney *U* test; Fig. 2b and Extended Data Fig. 5b). Next, we compared untreated HMF cancers to HMF tumors that had been exposed to anticancer treatment before the tumor biopsy collection. Pretreated HMF tumors showed a further significant increase (0.57, 95% CI: 0.50, 0.63, $P = 3.8 \times 10^{-3}$; Fig. 2b). A resampling analysis in which the number of samples per tumor type was equal between primary cancer cohort, untreated advanced cancer and treated advanced cancer cohort sets demonstrated that the ecDNA frequency increase following therapy exposure is independent of tumor type (Extended Data Fig. 5c). Grouping of HMF patients by the number of pretreatments demonstrated that the ecDNA frequency increase correlated with the number of therapies received (Fig. 2c and Extended Data Fig. 6a). We repeated this analysis in two tumor types with at least 20 samples per pretreatment group and observed the same trend in colorectal cancer, but not in breast cancer (Extended Data Fig. 6b). Further grouping of previously treated HMF patients by treatment class showed that chemotherapy demonstrates the strongest association with ecDNA frequency (Fig. 2d and

https://doi.org/10.1038/s41588-024-01949-7

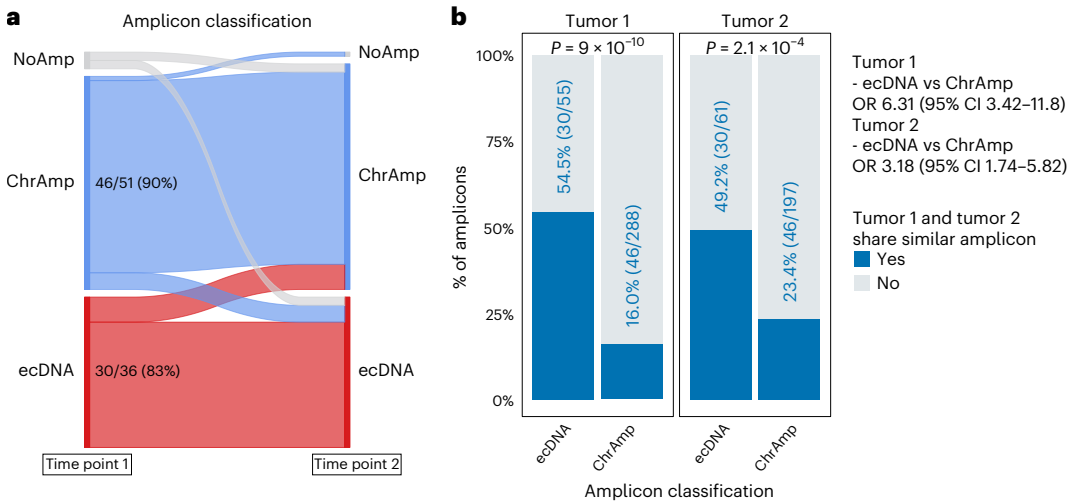


Fig. 3 | Longitudinal amplicon analysis. a, Sankey plot showing amplicon classification over time. Only amplicon pairs with statistically significant similarity were included ($n = 91$). Colors reflect amplicon classification, and numbers indicate the number of amplicons retained between two time points over all amplicons from the first tumor in the corresponding amplicon category. **b**, The fraction of ecDNA and ChrAmp amplicon pairs retained between the first and the second tumor. Numbers in parentheses indicate the numbers of first tumor amplicons also detected in the second tumor, over the number of all first tumor amplicons. *P* value was calculated using the chi-square test for tumors 1 and 2. OR, odds ratio.

Extended Data Fig. 6c). Tumors from patients treated with targeted therapy contained fewer ecDNAs compared to untreated tumors in the advanced cohort. Targeted therapies may specifically inhibit oncogenes carried on ecDNAs, which has been related to ecDNA genome reintegration as a mechanism of therapy resistance[28]. We evaluated whether pretreatment with a targeted inhibitor altered the ratio of oncogene target-carrying ecDNAs to chromosomal amplifications by comparing the observed ratio to a randomly sampled background distribution from comparable untreated cohorts. We found that the actual ratio was significantly higher compared to the background distribution, suggesting that treatment using inhibitors of oncogenes amplified on ecDNAs did not result in the formation of ChrAmps (Extended Data Fig. 6d).

To investigate whether different types of chemotherapy showed different associations with the number of ecDNAs, we categorized chemotherapy mechanisms into the following three types: antimetabolite, DNA damage agent and tubulin inhibitor. HMF patients pretreated with tubulin inhibitor had a higher ecDNA frequency (Extended Data Fig. 6e). The trend observed in the ecDNA counts mirrored that of the ChrAmp counts, which may indicate that antitubulin therapy results in genomic instability that leads to the formation of new amplicons (Extended Data Fig. 6e,f)[29,30]. These observations implicate newly acquired focal amplifications as a marker for therapy response and suggest that specific anticancer therapies may act as drivers of amplicon formation.

ecDNAs are preferentially preserved over time

Among patients whose tumors have been sequenced as part of TCGA and HMF, a subset ($n = 131$) was enrolled multiple times, resulting in WGS profiles from multiple time points[31]. The availability of longitudinal datasets provides an opportunity for evaluation of the stability and evolution of ecDNA structure. Time-separated whole-genome tumor sequences were also available through the GLASS consortium ($n = 100$)[21,32,33]. We constructed a cohort of 153 patients with multiple whole genomes passing quality filters (Extended Data Fig. 1d). The dataset includes 70 glioblastomas and gliomas, 18 prostate cancers, 16 breast cancers and 49 matched samples from other tumor types.

In total, 343 amplicons were detected at the first time point (T1), of which 55 amplicons were extrachromosomal. At time point 2 (T2), 258 amplicons were detected, including 61 ecDNAs. To determine how often amplicons were maintained over time, we determined amplicon similarity in a pair-wise fashion[23]. An amplicon similarity metric ranging from 0 to 1 was computed between two amplicons with overlapping territory based on shared breakpoints and genomic content. Specifically, 30 of 55 (54.5%) ecDNA and 46 of 288 (16%) ChrAmp T1 amplicons were found to match a T2 amplicon with a statistically significant similarity score. In the majority, amplicons classified as either ecDNA or ChrAmp maintained the amplicon class at T2, with 30 of 36 T1-ecDNA/T2-ecDNA amplicons and 46 of 51 T1-ChrAmp/T2-ChrAmp amplicons (Fig. 3a). Similarly, 82% of T1 samples classified as ecDNA/ChrAmp/NoAmp were assigned to the same class at T2 (Extended Data Fig. 7a). We evaluated the amplicon location and structure of five HMF-derived T1-ecDNA amplicons that were initially classified as ChrAmp at T2. Those ChrAmp amplicons were detected in tumors with tumor purity >0.7 and mean tumor genome sequence coverage >93×, substantiating that the amplicon classification was accurate. Genomic reintegration of ecDNA elements has been observed in response to treatment[28]. However, we did not detect sequence reads linking the T2-ChrAmp amplicons outside their original location of the genome (Extended Data Fig. 7b–f). We, therefore, suggest that the classification change from ecDNA to ChrAmp is not the result of reintegration but of clonal selection; that is, the ecDNA clone is dominant in the T1 tumor but has been outcompeted by a clone driven by a ChrAmp amplicon in T2.

At both time points, the fraction of ecDNA amplicons with a matching ecDNA amplicon in the reciprocal tumor was significantly higher compared to the fraction of matching ChrAmp amplicons, showing that ecDNA amplifications are more likely to be retained over time (Fig. 3b). Amplicon pairs did not show significant differences in amplicon complexity, amplicon copy number or amplicon size (Extended Data Fig. 8a–c).

Next, we evaluated clustered mutation event frequency, as we found higher rates of kataegis in ecDNAs from the advanced cancer cohort compared to the primary cancer cohort. Confirming our observations from the singleton cohorts, we found that the number of clustered mutation events was significantly higher in ecDNA compared to ChrAmp amplicons (Extended Data Fig. 8d). The fraction of amplicons containing one or more clustered mutation events was significantly higher in ecDNA as well as ChrAmp amplicons that were shared, compared to amplicons that were private to one of the two time points. This finding was true when counting clustered mutations at T1 as well

https://doi.org/10.1038/s41588-024-01949-7

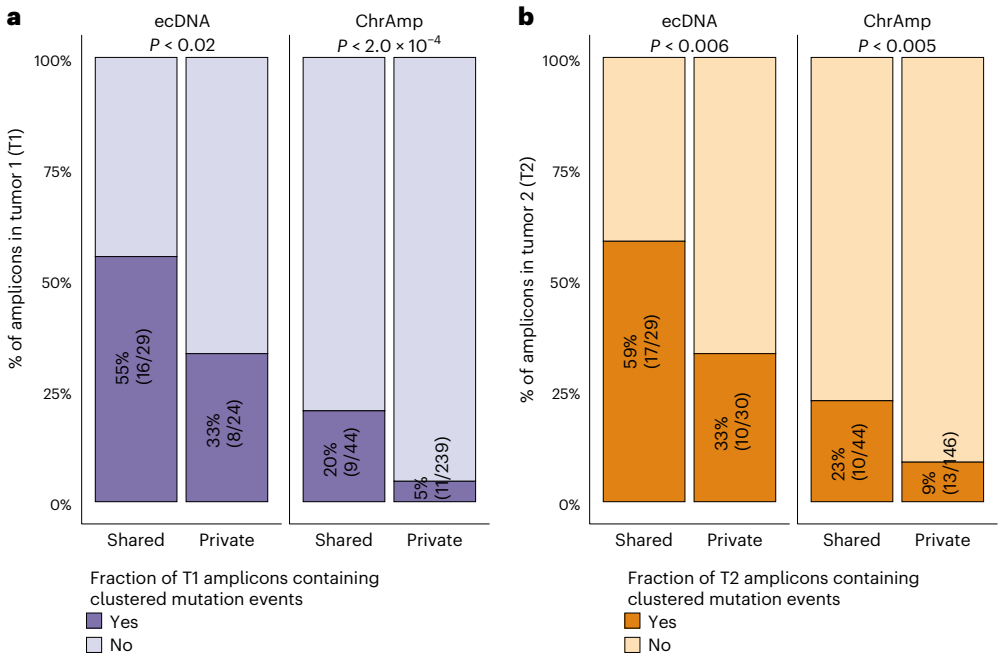


Fig. 4 | Clustered mutation events by amplicon category. a, The fraction and the number of ecDNA and ChrAmp amplicons with overlapping clustered mutation events in the T1 tumor. *P* values were computed using a binomial test (two-sided) with the fraction in the private category as a null probability for ecDNA and ChrAmp, respectively. **b**, The fraction and number of ecDNA and ChrAmp amplicons with overlapping clustered mutation events in the T2 tumor. *P* values were computed using a binomial test (two-sided) with the fraction in the private category as a null probability for ecDNA and ChrAmp, respectively.

as at T2 (Fig. 4a,b). Vice versa, T1 ecDNAs and T1 ChrAmps were more likely to be preserved at T2 when marked by a clustered mutation event (Extended Data Fig. 9a,b). Further separating amplicons by oncogene status suggested that these results are independent of whether an oncogene is present on the amplicon, while the analysis was limited by smaller numbers (Extended Data Fig. 9c,d).

We evaluated the variant allele fractions of clustered and nonclustered mutations on ecDNA and ChrAmp amplicons. Clustered mutations showed significantly higher variant allele fractions compared to nonclustered mutations at both T1 and T2 (Fig. 5a). There was no statistically significant difference in variant allele fraction between clustered mutations detected in private compared to shared ecDNAs. To complement this analysis and adjust for possible differences in tumor purity and ploidy, we inferred mutation cancer cell fractions. Mutations on shared ecDNAs showed significantly higher cancer cell fractions compared to mutations on private ecDNAs (Fig. 5b). Both shared and private T2 clustered mutation events were carried out at significantly higher cancer cell fractions compared to nonclustered mutations. Comparable patterns were observed among ChrAmp amplicons (Extended Data Fig. 10). Combined, the differences observed between variant allele and cancer cell fraction levels of shared and private ecDNAs and ChrAmps reflect that shared ecDNAs have undergone selection over a longer period of time. In addition, the higher variant allele and cancer cell fraction of clustered relative to nonclustered mutations suggest that clustered mutations generally occurred earlier in the amplicon lifetime.

Discussion

Activation of oncogenes through genomic amplification is a common event in cancer. TCGA and other -omic profiling efforts have provided a catalog of somatic alterations at diagnosis. Other initiatives, including the HMF, GLASS and tracking cancer evolution through therapy (TRACERx), are contributing to our understanding of how the molecular foundation of cancer diversifies over space and time[20,21,34]. By comparing data across different cohorts using conservative quality filters, we found that focal amplifications on ecDNA elements can be commonly detected in cancer. As described in the first half of this paper, the fraction of cancers significantly increased in metastatic and/or previously treated tumors. The penetrance of chromosomal focal amplifications also increased with tumor progression. The genomic landscape of cancer is under strong selection, and the increased amplicon frequency in advanced cancers suggests that the new formation of focal amplifications provides specific benefits to tumors postdiagnosis. In accordance with this observation, we observed an increase in the number of ecDNAs and ChrAmps per tumor following anticancer treatment, with the greatest gain associated with chemotherapy. Among different types of chemotherapy, tubulin inhibition via drugs such as paclitaxel and docetaxel provided the largest contribution to the increase in ecDNA and ChrAmps. This finding may warrant further investigation to understand whether tubulin inhibition drives amplicon formation and whether amplicon formation has a role in rendering tumor cells resistant to tubulin inhibition.

Surveillance of genomic integrity surveillance becomes increasingly error-prone as cancer progresses[35,36], and the resulting genomic instability may create opportunities for the genesis of ecDNA. In environments where cancer cells compete for resources such as oxygen and nutrients, or in response to the stress imposed by anticancer treatments and during metastasis, focal amplifications and ecDNAs in particular may provide opportunities for adaptation that afford cancer cells with higher proliferation rates. As we observed that ecDNAs were retained over time at higher rates compared to chromosomal amplicons, the uneven segregation of ecDNAs[7,9] likely contributes to their competitive advantage during the Darwinian process. Future studies of treatment resistance under controlled circumstances in model systems are needed to elucidate the mechanisms through which focal amplifications enhance untargeted therapy responses. In the second half of our paper, we presented evidence that a small subset of ecDNAs in our analysis were replaced by similar chromosomal amplicons at a later time point. Reintegration of ecDNAs near chromosome ends has been shown to occur following DNA damage[13,37]. However, for ecDNA reintegration to

https://doi.org/10.1038/s41588-024-01949-7

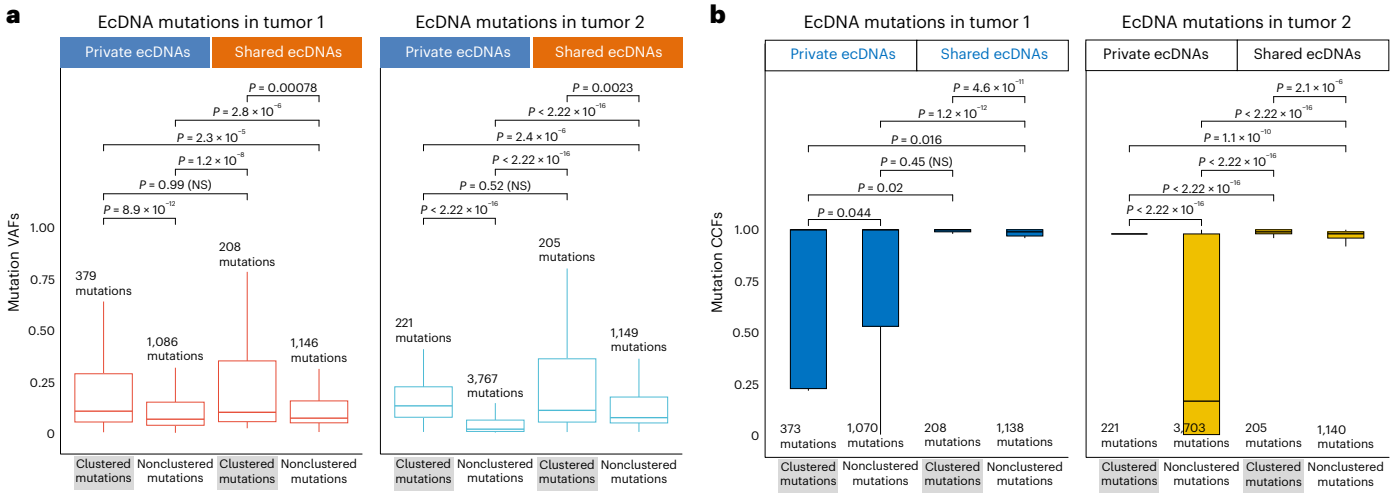


Fig. 5 | Variant allele fraction by mutational category. a,**b**, Comparison of (**a**) VAFs and (**b**) CCFs of different mutational categories detected on longitudinally shared or private ecDNA amplicons. Boxplots represent minimum (0th percentile), maximum (100th percentile), first and third quartiles and median with outliers excluded. *P* values were calculated using a two-sided Mann–Whitney *U* test. VAFs, variant allele fractions; CCFs, cancer cell fractions.

be detectable with sequencing, a single integration locus would have to be carried in a sufficient number of cells to overcome the sensitivity thresholds of sequencing, which would likely only occur if specific reintegration events underwent positive selection. Thus, the switching of ecDNA to ChrAmps that we observed is more likely to reflect the positive selection of pre-existing ChrAmps, rather than the reintegration of the ecDNA molecule. This is substantiated by the finding that these ChrAmps were detected at their original location in the genome, rather than near genome ends[37]. However, the precise delineation of chromosomal and extrachromosomal amplification structures in tumors where multiple subclones in parallel amplify the same genomic locus remains a challenge. Such amplicon heterogeneity may provide an orthogonal explanation for observations of amplicon class switching.

The short-read sequencing technology used to characterize cancer genomes in the cohorts analyzed here may pose limitations on the ability to detect amplicons with high sensitivity and characterize their structure, as well as the sensitivity to detect ecDNAs that have reintegrated into the genome. We aimed to address these limitations by imposing quality filters that accounted for tumor purity and genome coverage. However, studies of substantial tumor cohorts analyzed through long-read or optical mapping methods are needed to overcome these barriers. Such approaches may also be able to detect ecDNA reintegration.

Jointly, our results provide further support for the potential of developing therapeutic anticancer strategies targeting ecDNAs, implying that one effective strategy would be to combine blocking ecDNA formation with limiting ecDNA maintenance.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-024-01949-7.

References

1. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46 (2022).
2. Seyfried, T. N. & Huysentruyt, L. C. On the origin of cancer metastasis. *Crit. Rev. Oncog.* **18**, 43–73 (2013).
3. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* **185**, 563–575 (2022).
4. Martinez-Jimenez, F. et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
5. Albertson, D. G. Gene amplification in cancer. *Trends Genet.* **22**, 447–455 (2006).
6. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer* **19**, 283–288 (2019).
7. Yi, E., Chamorro Gonzalez, R., Henssen, A. G. & Verhaak, R. G. W. Extrachromosomal DNA amplifications in cancer. *Nat. Rev. Genet.* **23**, 760–771 (2022).
8. Kim, H. et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).
9. Yi, E. et al. Live-cell imaging shows uneven segregation of extrachromosomal DNA elements and transcriptionally active extrachromosomal DNA hubs in cancer. *Cancer Discov.* **12**, 468–483 (2021).
10. Barker, P. E., Drwinga, H. L., Hittelman, W. N. & Maddox, A. M. Double minutes replicate once during S phase of the cell cycle. *Exp. Cell Res.* **130**, 353–360 (1980).
11. deCarvalho, A. C. et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).
12. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
13. Koche, R. P. et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat. Genet.* **52**, 29–34 (2020).
14. Hung, K. L. et al. ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731–736 (2021).
15. Wu, S. et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699–703 (2019).
16. Shimizu, N., Itoh, N., Utiyama, H. & Wahl, G. M. Selective entrapment of extrachromosomally amplified DNA by nuclear budding and micronucleation during S phase. *J. Cell Biol.* **140**, 1307–1320 (1998).
17. Von Hoff, D. D. et al. Elimination of extrachromosomally amplified MYC genes from human tumor cells reduces their tumorigenicity. *Proc. Natl Acad. Sci. USA* **89**, 8165–8169 (1992).
18. Morton, A. R. et al. Functional enhancers shape extrachromosomal oncogene amplifications. *Cell* **179**, 1330–1341 (2019).

19. Helmsauer, K. et al. Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma. *Nat. Commun.* **11**, 5823 (2020).
20. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
21. Barthel, F. P. et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
22. Deshpande, V. et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 392 (2019).
23. Luebeck, J. et al. Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature* **616**, 798–805 (2023).
24. Bergstrom, E. N. et al. Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* **602**, 510–517 (2022).
25. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* **183**, 197–210 (2020).
26. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
27. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
28. Nathanson, D. A. et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* **343**, 72–76 (2014).
29. Scribano, C. M. et al. Chromosomal instability sensitizes patient breast tumors to multipolar divisions induced by paclitaxel. *Sci. Transl. Med.* **13**, eabd4811 (2021).
30. Crasta, K. et al. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).
31. Van de Haar, J. et al. Limited evolution of the actionable metastatic cancer genome under therapeutic pressure. *Nat. Med.* **27**, 1553–1563 (2021).
32. Varn, F. S. et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. *Cell* **185**, 2184–2199 (2022).
33. GLASS Consortium Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis consortium. *Neuro Oncol* **20**, 873–884 (2018).
34. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
35. Fitzgerald, D. M., Hastings, P. J. & Rosenberg, S. M. Stress-induced mutagenesis: implications in cancer and drug resistance. *Annu. Rev. Cancer Biol.* **1**, 119–140 (2017).
36. Tubbs, A. & Nussenzweig, A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168**, 644–656 (2017).
37. Shoshani, O. et al. Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**, 137–141 (2021).

[1]Department of Biopharmaceutical Convergence, School of Pharmacy, Sungkyunkwan University, Suwon-si, South Korea. [2]Department of Biohealth Regulatory Science, School of Pharmacy, Sungkyunkwan University, Suwon-si, South Korea. [3]Epigenome Dynamics Control Research Center, Sungkyunkwan University, Suwon-si, South Korea. [4]Department of Neurosurgery, Yale School of Medicine, New Haven, CT, USA. [5]Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [6]NORLUX Neuro-Oncology Laboratory, Department of Cancer Research, Luxembourg Institute of Health, Luxembourg, Luxembourg. [7]Department of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon-si, South Korea. [8]School of Biological Sciences, University of California at San Diego, La Jolla, CA, USA. [9]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine (FSTM), University of Luxembourg, Belvaux, Luxembourg. [10]Department of Neurosurgery, Cancer Research Institute and Ischemic/Hypoxic Disease Institute, Seoul National University College of Medicine, Seoul, South Korea. [11]Advanced Institutes of Convergence Technology, Seoul National University, Suwon-si, South Korea. [12]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA. [13]Halıcıoğlu Data Science Institute, University of California at San Diego, La Jolla, CA, USA. [14]Department of Neurosurgery, Amsterdam University Medical Centers/VUmc, Amsterdam, the Netherlands. [15]These authors contributed equally: Hoon Kim, Soyeon Kim. [16]These authors jointly supervised this work: Hoon Kim, G. W. Verhaak. ✉e-mail: wisekh@skku.edu; roel.verhaak@yale.edu

Methods

Ethical approval

This study reanalyzes data generated from previously published studies (TCGA, ICGC, HMF and GLASS) that complied with ethical regulations.

Patient cohort

The HMF cohort consists of metastatic tumor samples obtained after local or systemic treatment and as part of the CPCT-02 (NCT01855477) and DRUP (NCT02925234) clinical trials. Patients treated for a wide range of tumor-type diagnoses at various hospitals across the Netherlands were enrolled in the trials. Biopsy specimens were sequenced at the core facilities of the HMF. WGS was performed for each sample according to standardized protocols. Detailed information on sequence platforms, capture kits and read length has been outlined in the HMF marker paper[20]. Data access approval was granted to H.K. as well as R.G.W.V. WGS CRAM files and PURity & Ploidy Estimator (PURPLE[20])-inferred copy-number segment files were accessible through a Google Cloud Platform. Mutation VCF files and associated metadata were downloaded from the HMF Database (https://database.hartwigmedicalfoundation.nl). In total, the HMF database included WGS data from 4,513 tumor biopsies (after excluding patients with insufficient informed consent).

WGS datasets from the GLASS consortium were collected and preprocessed as previously reported[21,32]. Mutation VCF files and associated metadata were downloaded from www.synapse.org/glass.

WGS datasets from TCGA were accessed through the Institute for Systems Biology Cancer Genomics Cloud (ISB-CGC; https://isb-cgc.appspot.com/), which provides a cloud-based platform for TCGA data analysis. The processed (hg19) and clinical data were available at the Genomic Data Commons (https://portal.gdc.cancer.gov) and the PancanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas).

WGS datasets from ICGC were processed on the Amazon Web Services Cloud. The associated metadata were obtained from the ICGC data portal at https://dcc.icgc.org/.

Longitudinal sample pairs of glioma and glioblastoma tumors were also collected from the Centre Hospitalier de Luxembourg (CHL, Neurosurgical Department) from patients who had given their informed consent. The study received official approval from the National Committee for Ethics in Research (CNER) Luxembourg, under protocol 201201/06. Additional longitudinal sample pairs of glioma and glioblastoma tumors were collected from the Department of Neurosurgery, Seoul National University Hospital. It was approved by the Institutional Review Board of Seoul National University Hospital (approval H-2004-049-1116), and all patients provided signed informed consent accordingly.

Collecting tumor stage information

We collected tumor stage information for TCGA (Genomic Data Commons PanCancer portal: https://gdc.cancer.gov/about-data/publications/pancanatlas), Pan-Cancer Analysis of Whole Genomes (PCAWG; ICGC portal: http://dcc.icgc.org/releases/PCAWG/) and HMF[4]. For our analysis, we simplified the original complex tumor stages into stages 1, 2, 3 and 4 by assigning stage 1 to those originally annotated as 1 (A/B), I (A/B) and T1N0M0; stage 2 to 2 (A/B), II (A/B), T0N1M0, T1N1M0, T2N0M0, T2N1M0 and T3N0M0; stage 3 to 3 (A/B/C), III (A/B/C), T0N2M0, T1N2M0, T2N2M0, T3N1M0, T3N2M0, T4(Any N)M0 and (Any T)N3M0; stage 4 to 4, IV, (Any T)(Any N)M1, (Any T)(Any N)M2 and (Any T)(Any N)M3. Nonstage four samples with incomplete TNM stage (including 'X') annotation were excluded, and all patients from the HMF cohort were considered as stage IV cancer.

AmpliconArchitect

AmpliconArchitect (part of AmpliconSuite-pipeline, v.0.1344.2) was run using default settings. This includes BAM file downsampling to 10x coverage before detection of seed regions, to normalize sequencing depth between samples. In a mixed cancer-type WGS cohort of 133 samples, running AmpliconArchitect with or without downsampling did not significantly alter the number of ecDNAs detected. AmpliconArchitect was run using the maximum wall time set to 72 h per sample in Google Cloud and 2 weeks in Amazon Cloud (https://github.com/AmpliconSuite/AmpliconSuite-pipeline). Candidate seed regions for inputs to AmpliconArchitect were identified with AmpliconSuite-pipeline.py, which uses CNVkit[38] for detecting DNA copy-number alterations and were defined as at least 50 kb in length and a minimum of 4.5 copy numbers. Reconstruction of amplicon structures is based on the full and not a downsampled BAM and not affected by downsampling. We evaluated seed region count and did not observe significant associations between seed region count and tumor coverage or tumor purity bins. A higher number of seed regions positively correlated with the number of ecDNAs detected and showed similar trends in both primary and advanced cancer cohorts. We further examined the association between the number of ecDNA amplicons detected and the number of candidate seed regions, as well as the size of seed regions. We observed comparable degrees of positive correlation between primary and advanced cohorts; that is, the number of seeding regions is related to the number of ecDNA detected, but they do not disproportionately affect ecDNA frequency in the primary and advanced cohorts. For analysis of longitudinally paired samples, the candidate seed regions identified from different tumors in the same patient were merged into an identical set of candidate seed regions for those tumors in the patient. AmpliconClassifier (v0.4.11) was invoked from AmpliconSuite-pipeline to predict the class of focal amplification and refine gene coordinates involved in the specific focal amplifications.

Amplicon complexity

Amplicon complexity was calculated using AmpliconClassifier Amplicon complexity scores, as previously reported in ref. 23. Scores were computed for each focal amplification using the AmpliconArchitect cycles file, which encodes paths identified by AmpliconArchitect in the copy-number-aware AmpliconArchitect breakpoint graph explaining the observed changes in copy number. The complexity score takes the distribution of copy-number flow values assigned to each genome path of a specific focal amplification type and computes a vector, which represents the fraction of the total copy number captured by each path, weighted by the length of the path. The score also incorporates a residual, which measures the weighted copy-number fraction after the first 80% explained, if any is still remaining (for example, no residual would remain if one genome path could explain all the copy numbers). The amplicon complexity score function then combines the entropies of the residual, nonresidual and the total number of genome segments in the focal amplification, with a high score indicating an amplicon with a more complex structure than an amplicon with a low score.

Amplicon similarity

Amplicon similarity score was computed to quantify the similarity between genomically overlapping amplicons from T1 and T2 tumors from identical patients, implemented with amplicon_similarity.py script, available in AmpliconClassifier (v.0.4.11, part of the AmpliconSuite, v.0.1344.2)[8]. Following the identification of T1–T2 amplicon pairs with overlapping genomic regions, an amplicon similarity score was calculated using shared breakpoints and shared genomic content. The similarity score was compared against similarity scores from unrelated overlapping amplicon distributions to compute a P value for the similarity score. Amplicon pairs with P values < 0.05 were included in our analysis as shared events.

Detection of clustered mutations

SigProfilerSimulator (v.1.1.4)[39] was first applied to quality-filtered, single-nucleotide variant-only VCF files to determine the intramutational

distance cutoff for each sample to only detect mutation clusters that were not likely to occur by chance. Each sample was simulated 100 times in the pentanucleotide context (the ±2 bp sequence context) while maintaining the same mutational burden per chromosome and preserving the transcriptional strand bias. SigProfilerClusters (v.1.0.11)[40] was then used to subclassify clustered mutations while performing a genome-wide mutational density correction. A window size of 1 Mb was used for correcting intramutational distances based on mutational density, and mutation variant allele frequencies were considered when subclassifying clustered mutations. From SigProfilerClusters output, kataegis mutations having an identical group number were considered as a single clustered event. Each clustered event was defined as ecDNA-overlapping kataegis if overlapped with ecDNA regions and ChrAmp-overlapping kataegis if overlapped with ChrAmp regions. Only the samples having the available mutation files for which the clustered mutation calling was successful were included in this analysis (single time point analysis—2,454 (58 failed, 97 no mutation file) of 2,609 PCAWG samples and 4,136 (34 failed) of 4,170 HMF samples; multitime point analysis —248 (2 failed) of 250 HMF samples and 181 (1 failed, 18 no mutation file) of 200 GLASS samples). HMF mutation files in the form of VCF were provided by the HMF, TCGA–ICGC mutation files were obtained from https://dcc.icgc.org/ in the form of MAF and GLASS mutation files were from www.synapse.org/glass.

Determining the number of pretreatments

Each entry of prebiopsy drugs annotation provided by the HMF consists of a patient identifier, treatment start date, end date, name of the drug, type of the drug and the drug mechanism. After filtering out drug treatment entries that occurred before the sample biopsy date, the number of unique entries for a patient was defined as the number of pretreatments the patient had received. The treatment annotation provided by the HMF included a drug classification into broad categories including chemotherapy, hormonal therapy and targeted therapy. We further subdivided chemotherapy drug treatments into the following four categories: (1) antmetabolite, (2) DNA damage, (3) tubulin inhibitor and (4) other, based on the literature review. A detailed classification of drugs by mechanism of action and associated references is provided in Supplementary Table 2.

Estimating cancer cell fractions of mutations

The cancer cell fractions of single-nucleotide variant mutations for HMF and GLASS multitime point samples whose mutation, copy number and tumor purity are available were computed by PyClone-VI (v.0.1.2) with default parameters. Mutations on sex chromosomes were excluded. Mutation, copy number and purity files for HMF samples were provided by the HMF, and the files for GLASS samples were from www.synapse.org/glass.

Statistical analysis

All data analyses were conducted in R (v.4.1.2) and Python (v.3.9.13). Statistical tests were not adjusted for multiple comparisons.

Reporting summary

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

Data availability

WGS from TCGA were accessed through the database of Genotypes and Phenotypes (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) under accession ID phs000178.v11.p8 (TCGA). WGS data from PCAWG/ICGC were downloaded from the ICGC at https://dcc.icgc.org/ (Data Access Compliance Office application DACO-753). The WGS and associated clinical data used in this study were made available by the HMF and were accessed under a license agreement (HMF DR-057 v.3.0). Data access can be obtained by completing a data request form. The form and detailed application procedures can be found at https://www.hartwigmedicalfoundation.nl/applying-for-data/. Processed sequencing data from the GLASS project used in this study are available on Synapse at https://www.synapse.org/glass. AmpliconSuite output files for TCGA are available at https://ampliconrepository.org/project/655bda68bba7c92509522479. AmpliconSuite output files for PCAWG are available at https://ampliconrepository.org/project/655c060abba7c925095555da. AmpliconSuite output files for GLASS are available at https://ampliconrepository.org.

Code availability

The code used for analysis has been deposited at https://github.com/hoonbiolab/panecmanuscript2024.

References

38. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
39. Bergstrom, E. N., Barnes, M., Martincorena, I. & Alexandrov, L. B. Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics* **21**, 438 (2020).
40. Bergstrom, E. N., Kundu, M., Tbeileh, N. & Alexandrov, L. B. Examining clustered somatic mutations with SigProfilerClusters. *Bioinformatics* **38**, 3470–3473 (2022).

Acknowledgements

This publication and the underlying study have been made possible partly on the basis of the data that the HMF and the Center of Personalized Cancer Treatment have made available to the study. We thank the research information technology team at the Jackson Laboratory for their support in setting up cloud-based analyses. We are grateful to the Neurosurgery Department of the CHL and the Clinical and Epidemiological Investigation Center of the LIH for support in tumor collection (www.precision-pdx.lu). Results published in this paper are in whole or in part based on data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and the ICGC (https://icgc.org/). Analysis of the TCGA and ICGC datasets was made possible through the ISB-CGC and the Amazon Web Services Cloud, respectively.
This work was delivered as part of the eDyNAmiC team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCATF-2021/100012; CGCATF-2021/100016 to R.G.W.V.; and CGCATF-2021/100025 to V.B. and J.L.) and the National Cancer Institute (OT2CA278688; OT2CA278649 to R.G.W.V.; and OT2CA278635 to V.B. and J.L.). This work was also supported by the National Institutes of Health (grants R01 CA237208, R21 NS114873 and R33 CA236681 to R.G.W.V.) and Cancer Center Support Grant (P30 CA034196, U24CA264379 and R01GM114362 to V.B.); the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT; NRF-2019R1A5A2027340 and NRF-2022M3C1A3092022 to H.K.), the Korea Health Industry Development Institute (KHIDI) grant funded by Ministry of Health & Welfare (HI19C1348 to S. Kim) and the Luxembourg National Research Fund (FNR; C20/BM/14646004/GLASSLUX to A.L., A.G. and S.P.N.).

Author contributions

H.K. and R.G.W.V. conceptualized the project. H.K., S. Kim, T.W., J.K., H.Y.L., J.L. and R.G.W.V. developed the methodology. H.K. and R.G.W.V. conducted the investigation. H.K., S. Kim, E.Y., Y.N. and S.A. handled visualization. H.K., S. Kim, R.G.W.V., K.C.J., S. Kang, J.K., H.Y.L., V.B., J.L., A.L., A.G., S.P.N., H.C., H.E.M. and S.H.P. curated the data and conducted the analysis. H.K., R.G.W.V. and S.P.N. secured funding. H.K. and R.G.W.V. managed project administration and provided supervision. H.K., S. Kim and R.G.W.V. did the writing. All authors have read and agreed to the contents of this manuscript.

Competing interests

R.G.W.V. is a cofounder of, holds equity in and has received research funds from Boundless Bio. H.K. has received research funds from JW Pharmaceutical. J.L. receives compensation as a part-time consultant for Boundless Bio. V.B. is a cofounder, paid consultant and Scientific Advisory Board member, and has an equity interest in Boundless Bio and Abterra. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies. All other authors declare no competing interests.

Additional information

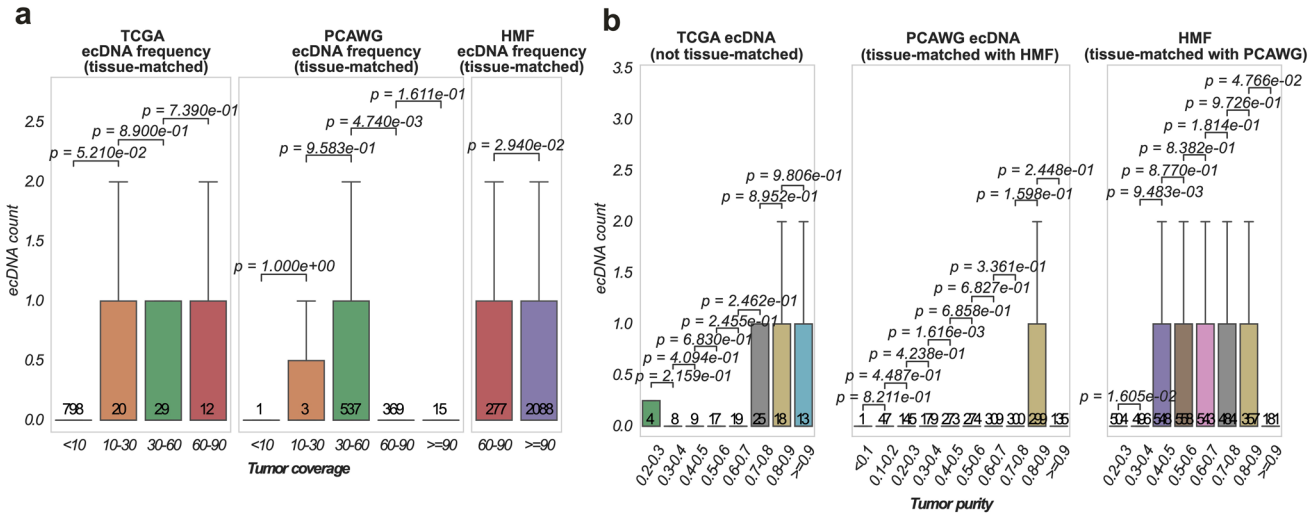
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01949-7.

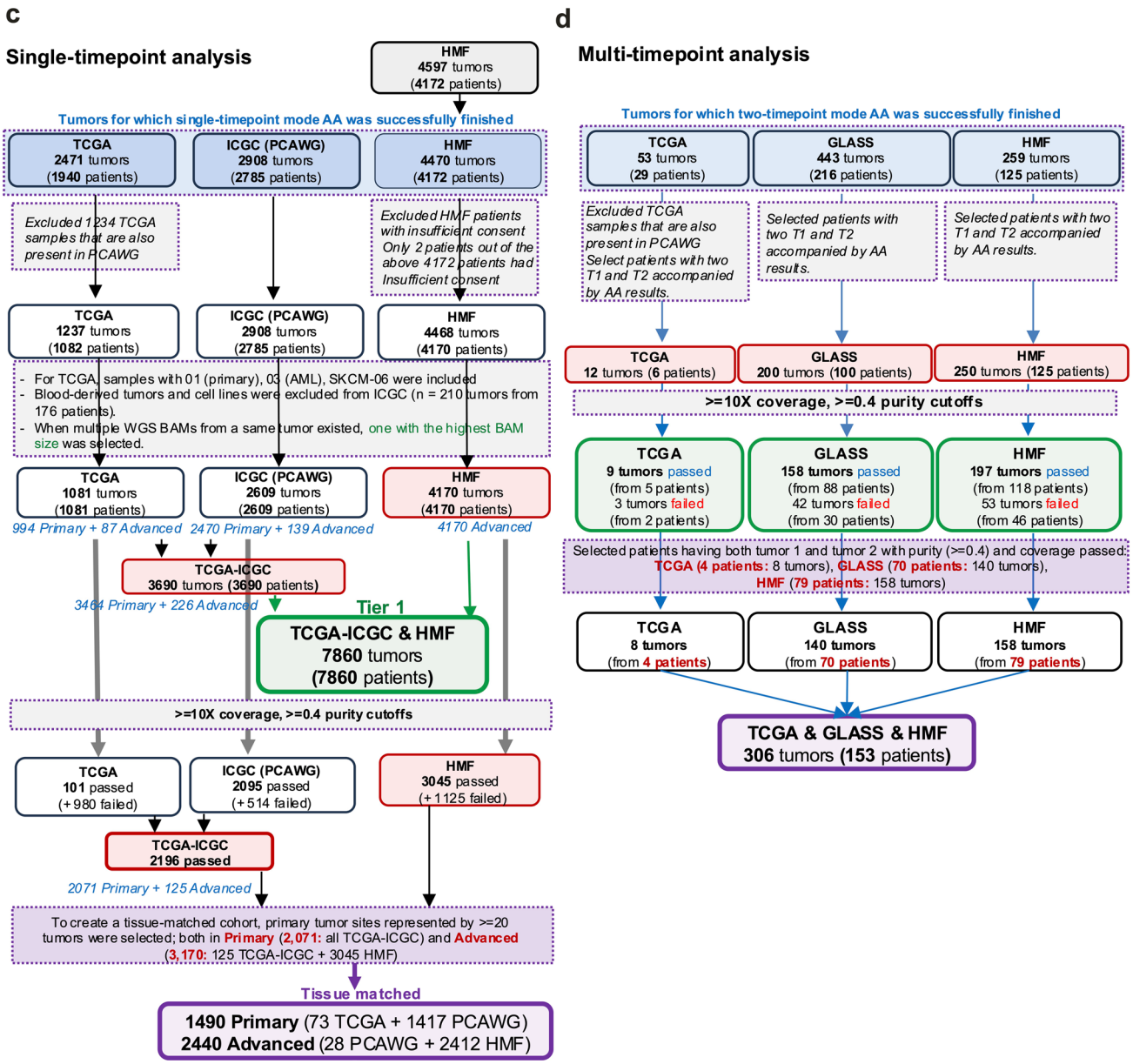
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01949-7.

Correspondence and requests for materials should be addressed to Hoon Kim or Roel G. W. Verhaak.

Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01949-7



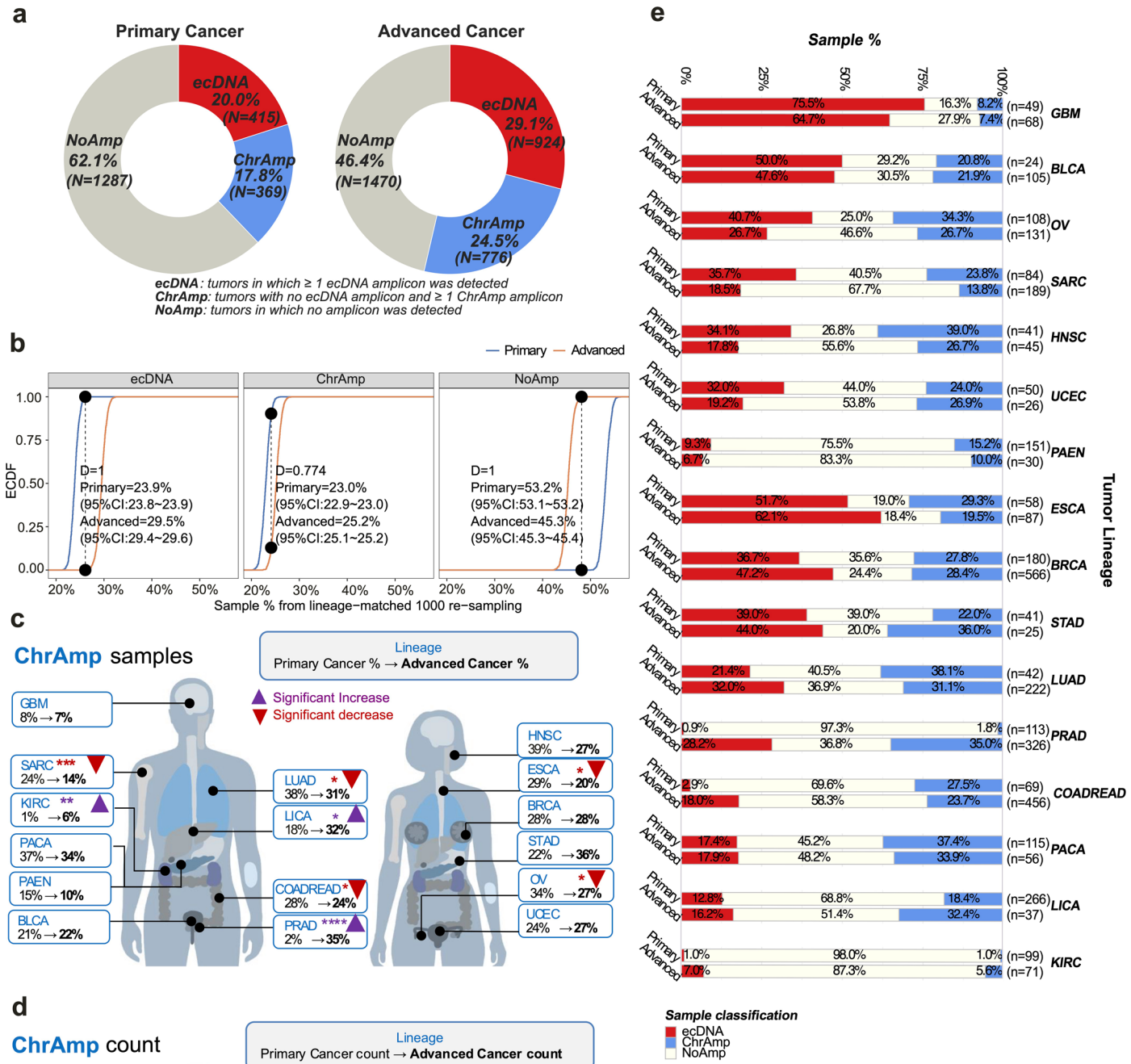


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-024-01949-7

Extended Data Fig. 1 | Overview of sample selection criteria. a, Comparison of extrachromosomal DNA (ecDNA) count by cohort and average sequence coverage. *P*-values are derived from a two-sided Mann–Whitney U test. Tissues are matched across the Cancer Genome Atlas (TCGA), the Pan-Cancer Analysis of Whole Genomes (PCAWG) and the Hartwig Medical Foundation (HMF; at least 20 samples in each cohort). Numbers on the bar indicate the number of samples. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median with outliers not shown. **b**, Comparison of ecDNA count by cohort and tumor purity bin for samples whose coverage is higher or equal to 10×. *P*-values are derived from a two-sided Mann–Whitney U test. TCGA includes all samples above the coverage cutoff. Tissues were only matched between PCAWG and HMF (at least 20 samples in both) because the TCGA sample size after coverage filtering was too small. Numbers on the bar indicate the sample number. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median with outliers not shown. **c**, Cohort and sample selection overview for single time point analysis. **d**, Cohort and sample selection overview for multitime point analysis. Abbreviations are defined as follows: AA, AmpliconArchitect tool; ICGC, International Cancer Genome Consortium; AML, acute myeloid leukemia; SKCM, skin cutaneous melanoma; T1, first time point tumor; T2, second time point tumor; GLASS, the Glioma Longitudinal Analysis Consortium.

https://doi.org/10.1038/s41588-024-01949-7

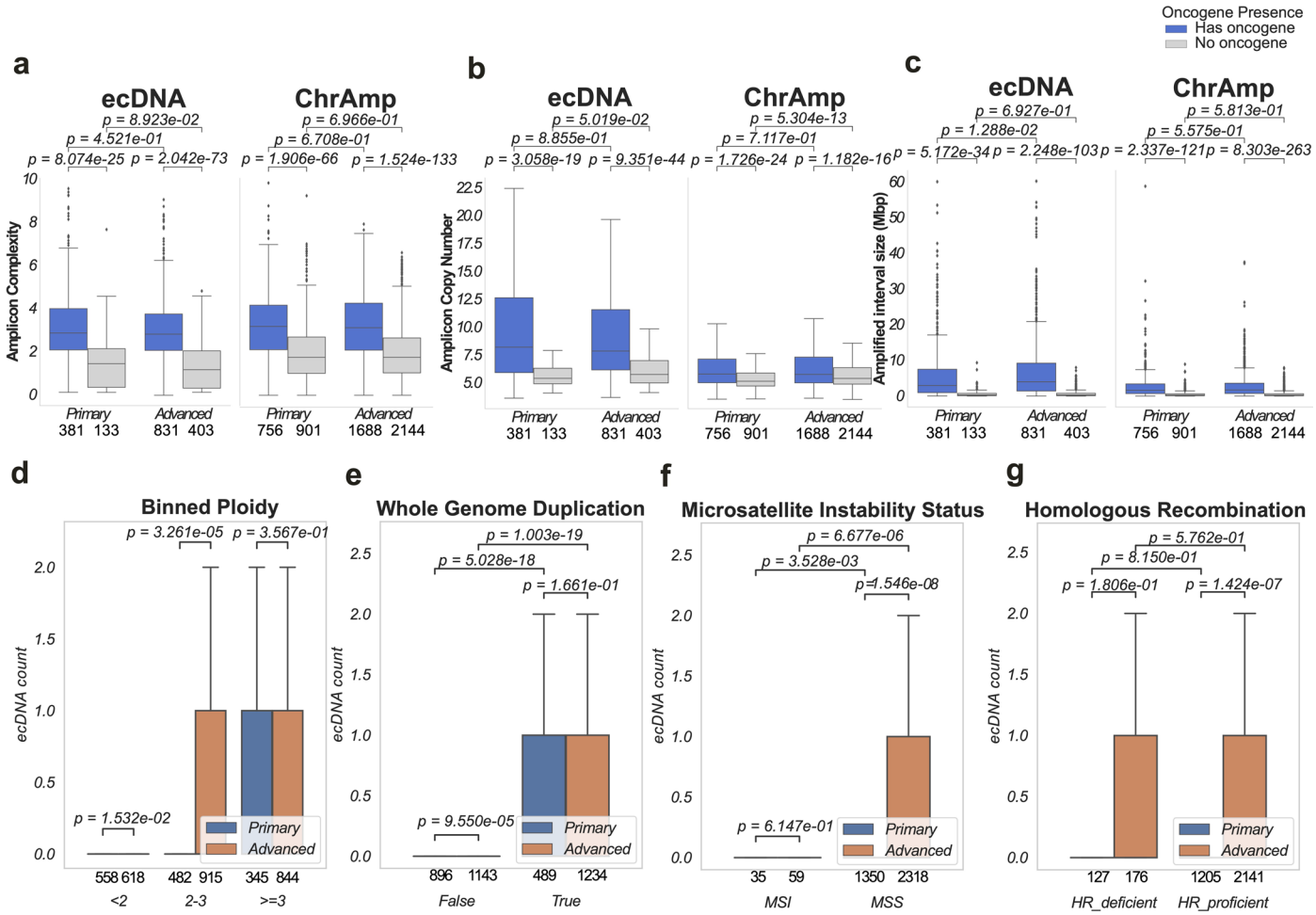


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-024-01949-7

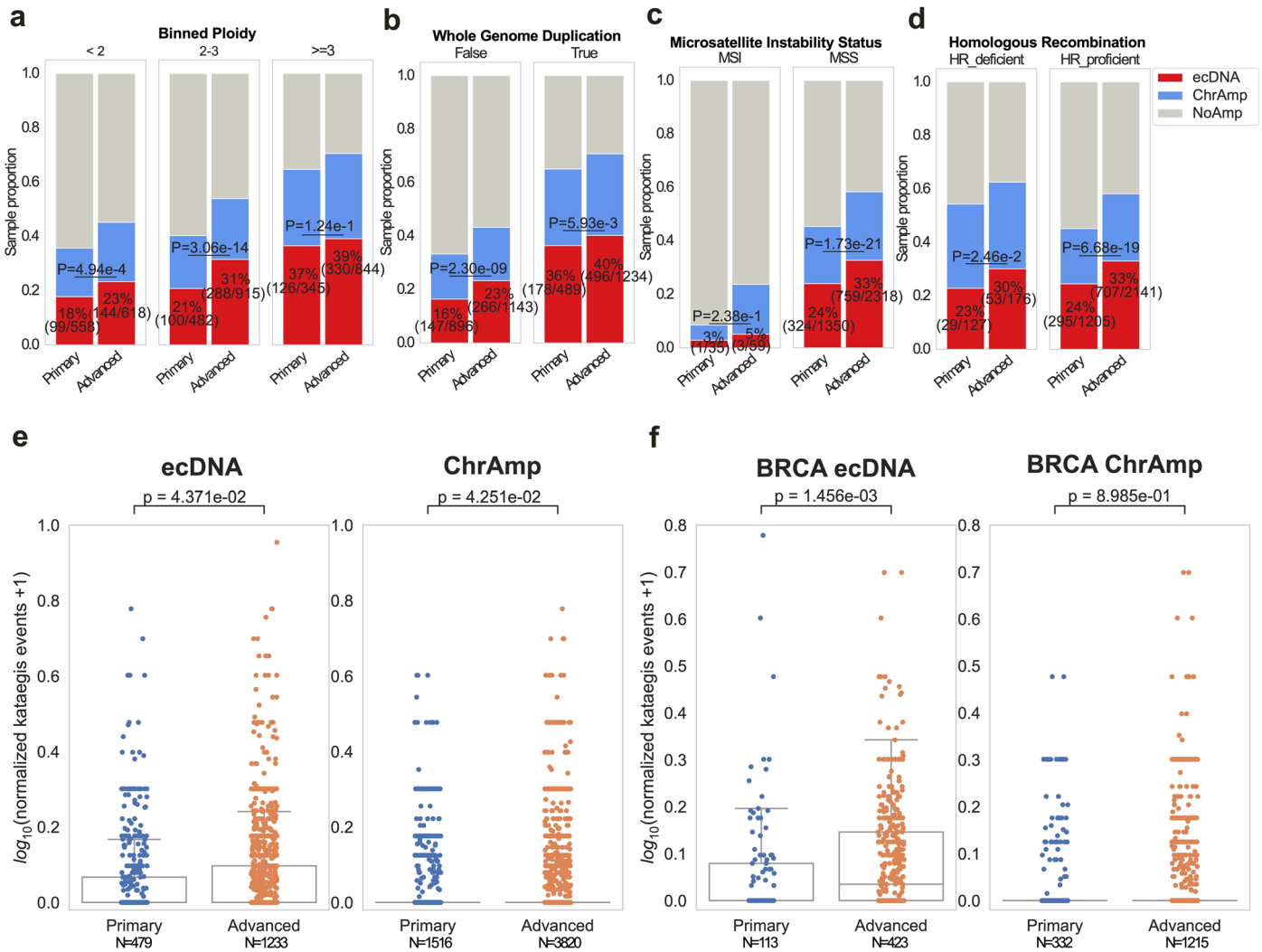
Extended Data Fig. 2 | Additional data to sample and amplicon classification. **a**, Overview of sample classification for the 2,071 primary and 3,170 advanced patients whose tumor sequencings are above purity and coverage cutoff, including all tumor types. Numbers in parentheses indicate number of tumor samples. **b**, Resampling analysis with replacement was repeated 1,000 times while maintaining sample count per tumor-type identical between primary cancer and advanced cancer cohorts in each resampled dataset to compare classification distributions shows a significant increase in the number of samples classified as ecDNA and ChrAmp, respectively, in the advanced cancer cohort, independent of tumor-type distribution. Empirical cumulative distributions (ECDF) of sample classification percentage using 1,000 re-sampled datasets. D represents Kolmogorov–Smirnov statistic. **c**,**d**, Percentage of ChrAmp samples (**c**) and the average number of distinct ChrAmp amplicons per sample (**d**) in primary and advanced cancer cohorts, showing tumor lineage represented by at least 20 tumors in both cohorts. P-values were computed using a one-sided binomial test with the ChrAmp-carrying tumor fraction in the primary cancer cohort as a null probability in **c** and using a one-sided Mann–Whitney U test in **d**. Not significant unless noted otherwise. Asterisks indicate level of significance: *1.00e−02 < p ≤ 5.00e−02; **1.00e−03 < p ≤ 1.00e−02; ***1.00e−04 < p ≤ 1.00e-03; ****p ≤ 1.00e−04. **e**, Distribution of primary and advanced sample classification stratified by tumor lineages each of which includes at least 20 tumors. Numbers in parentheses indicate the number of ecDNA samples and the total number of samples of that lineage.

https://doi.org/10.1038/s41588-024-01949-7



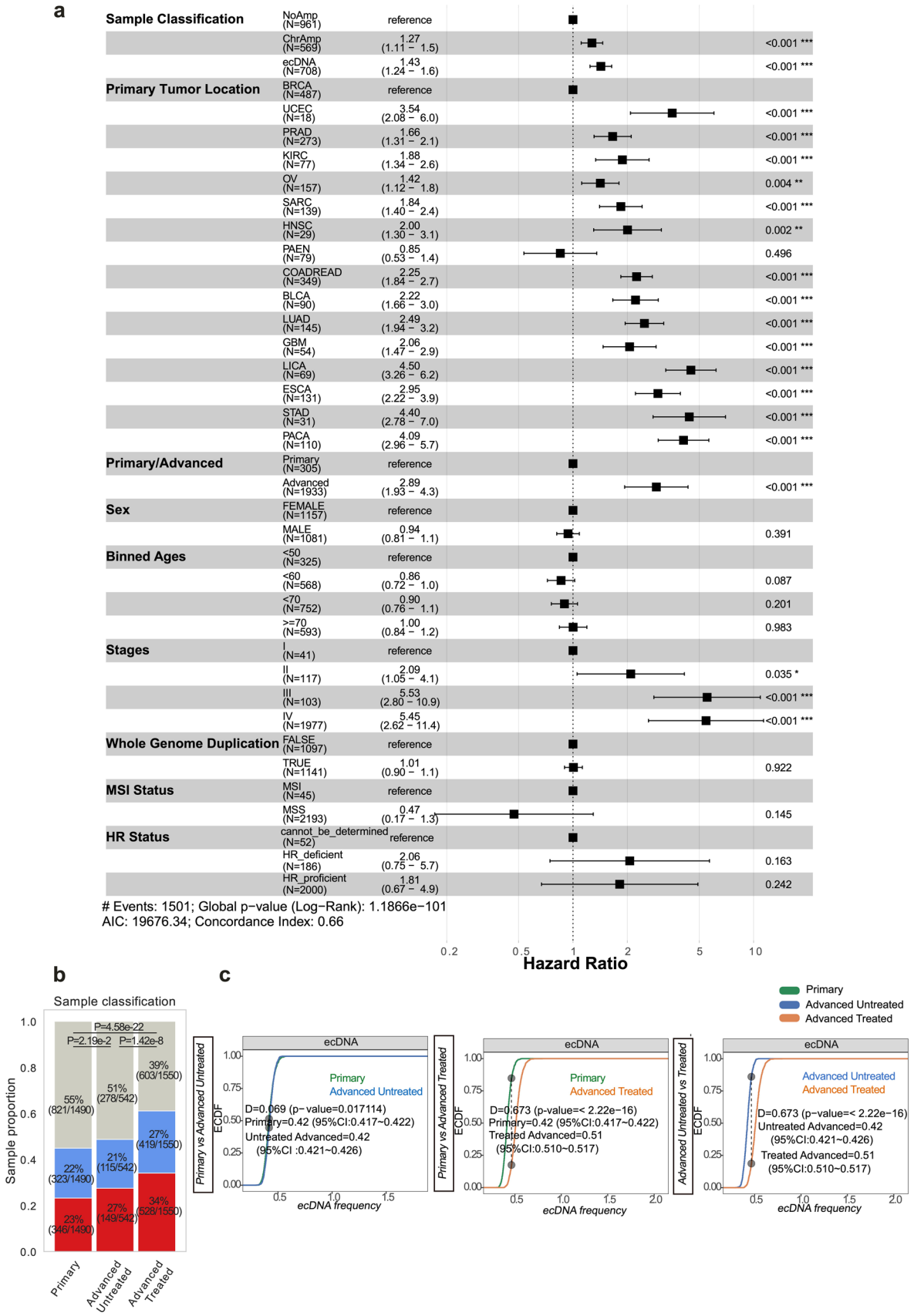
Extended Data Fig. 3 | Amplicon properties by amplicon class and oncogene presence. a, Box plot showing amplicon complexity. **b**, Box plot showing amplicon DNA copy number. **c**, Box plot showing amplicon size. Numbers indicate number of amplicons. P-values were computed using a two-sided Mann–Whitney U test. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median. For Extended Data Fig. 3b, outliers are not plotted. **d**–**g**, Comparison of ecDNA amplicon count per patient between primary and advanced cohorts when further grouping patients according to measures of genomic instability, (**d**) including binned ploidy; (**e**) whole-genome duplication status; (**f**) microsatellite instability status; and (**g**) homologous recombination (HR) status. Numbers indicate number of patients. P-values were computed using a two-sided Mann–Whitney U test. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median with outliers not shown. MSS, microsatellite stable; MSI, microsatellite Instable.

https://doi.org/10.1038/s41588-024-01949-7



Extended Data Fig. 4 | Comparison of ecDNA patient fractions when further grouping patients according to measures of genomic instability. a–**d**, Binned ploidy (**a**), whole-genome duplication status (**b**), microsatellite instability status (**c**) and homologous recombination status (**d**). Numbers in parentheses represent number of patients carrying ecDNA over all patients in the category. P values were calculated using a two-sided binomial with the ecDNA-carrying tumor category in the primary cohort as a null probability. **e**, Number of kataegis events normalized by the number of intervals present on ecDNA or ChrAmp amplicons between primary cancer and advanced cancer cohorts. Plots show log plus one transformed value on the y-axis. P values were calculated using a two-sided Mann–Whitney U test. **f**, Same but breast cancer samples only. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median.

https://doi.org/10.1038/s41588-024-01949-7

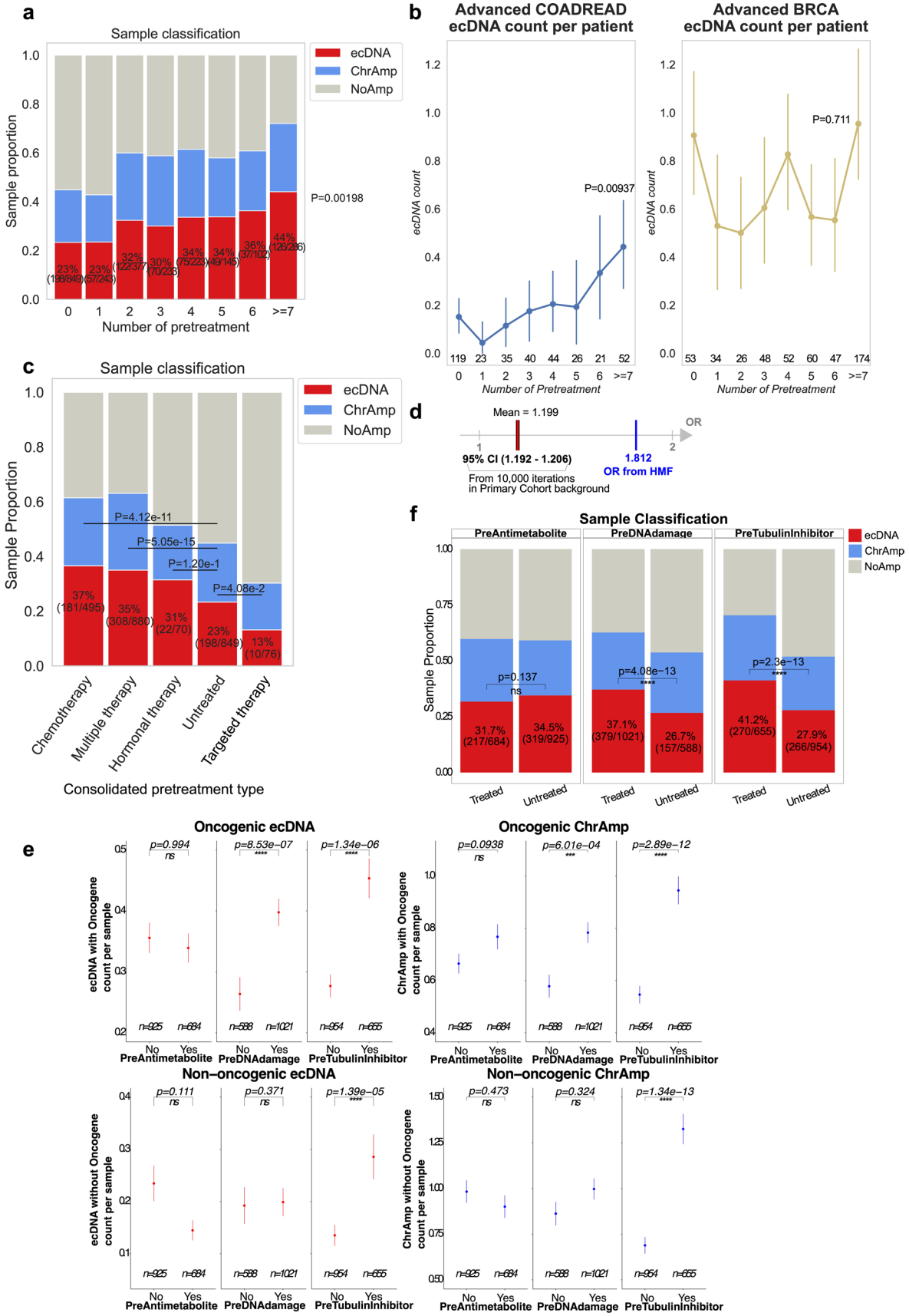


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41588-024-01949-7

Extended Data Fig. 5 | Additional data to clinical associations. a, Multivariate Cox proportional hazards model, incorporating primary tumor locations, sex, age, whole-genome doubling status, microsatellite instability (MSI) status, homologous recombination (HR) status and tumor stage in primary and advanced cancer cohorts, showing that extrachromosomal DNA amplification resulted in the highest hazard ratio. The error bars represent the 95% confidence intervals of the hazard ratios. Asterisks indicate level of significance: *$1.00 \times 10^{-2} < p \leq 5.00 \times 10^{-2}$; **$1.00 \times 10^{-3} < p \leq 1.00 \times 10^{-2}$; ***$1.00 \times 10^{-4} < p \leq 1.00 \times 10^{-3}$. **b**, Distribution of primary, advanced untreated and advanced treated cohorts into ecDNA/ChrAmp/NoAmp categories. All tumors with available pretreatment information were included in the analysis. Y-axis represents category fractions. Numbers indicate patient counts. P-values were computed using a two-sided binomial test with the ecDNA-carrying tumor fraction in the primary cancer cohort as a null probability when comparing primary vs advanced untreated/treated and that in the advanced untreated cohort as a null probability when comparing advanced untreated vs advanced treated. **c**, Resampling analysis with replacement was repeated 1,000 times while maintaining sample count per tumor-type identical between primary cancer and advanced cancer untreated and advanced cancer treated cohorts, in each resampled dataset, to compare classification distributions. Empirical cumulative distributions of sample classification percentage using 1,000 re-sampled datasets. D represents Kolmogorov–Smirnov statistic (two-sided).

https://doi.org/10.1038/s41588-024-01949-7

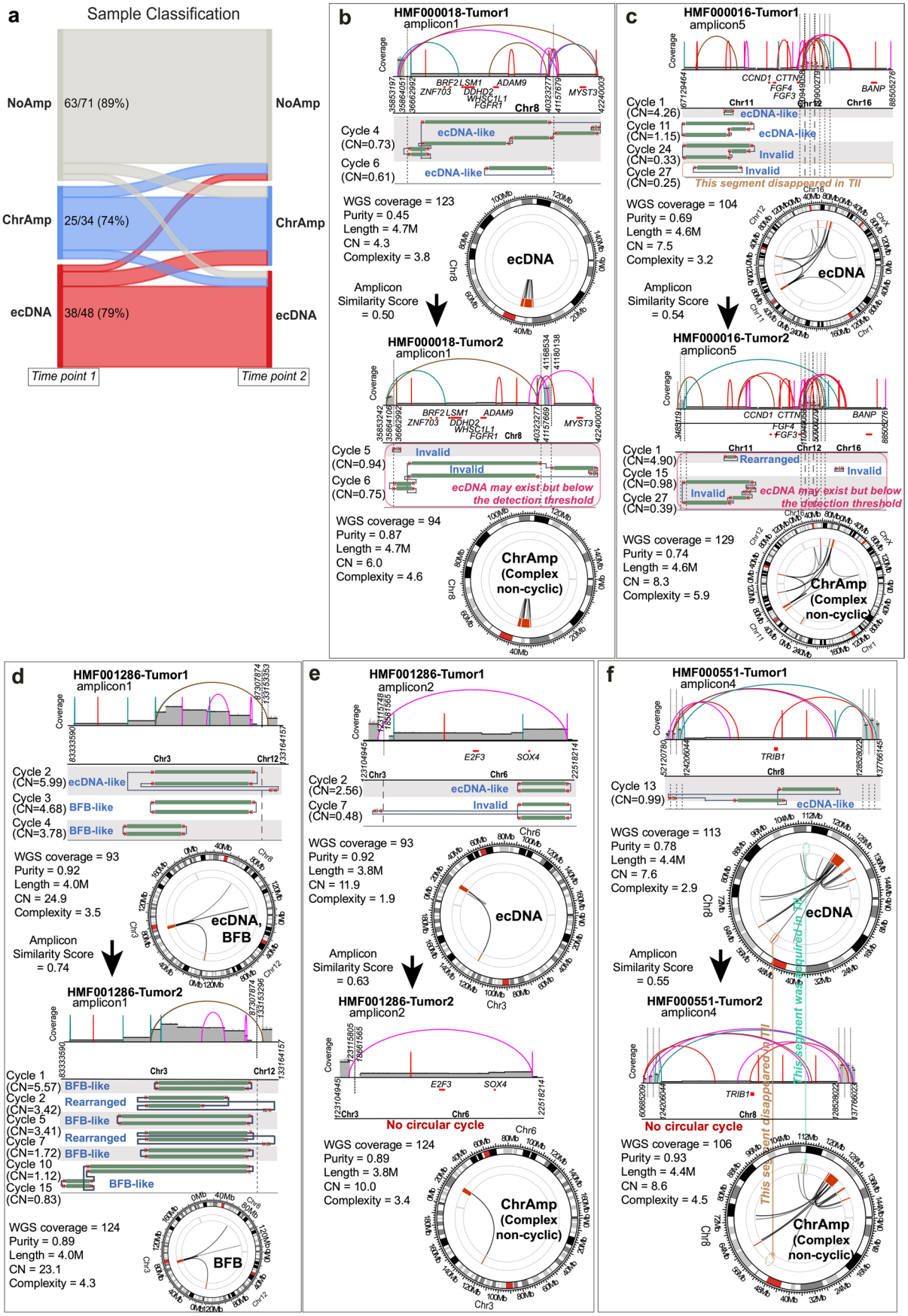


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-024-01949-7

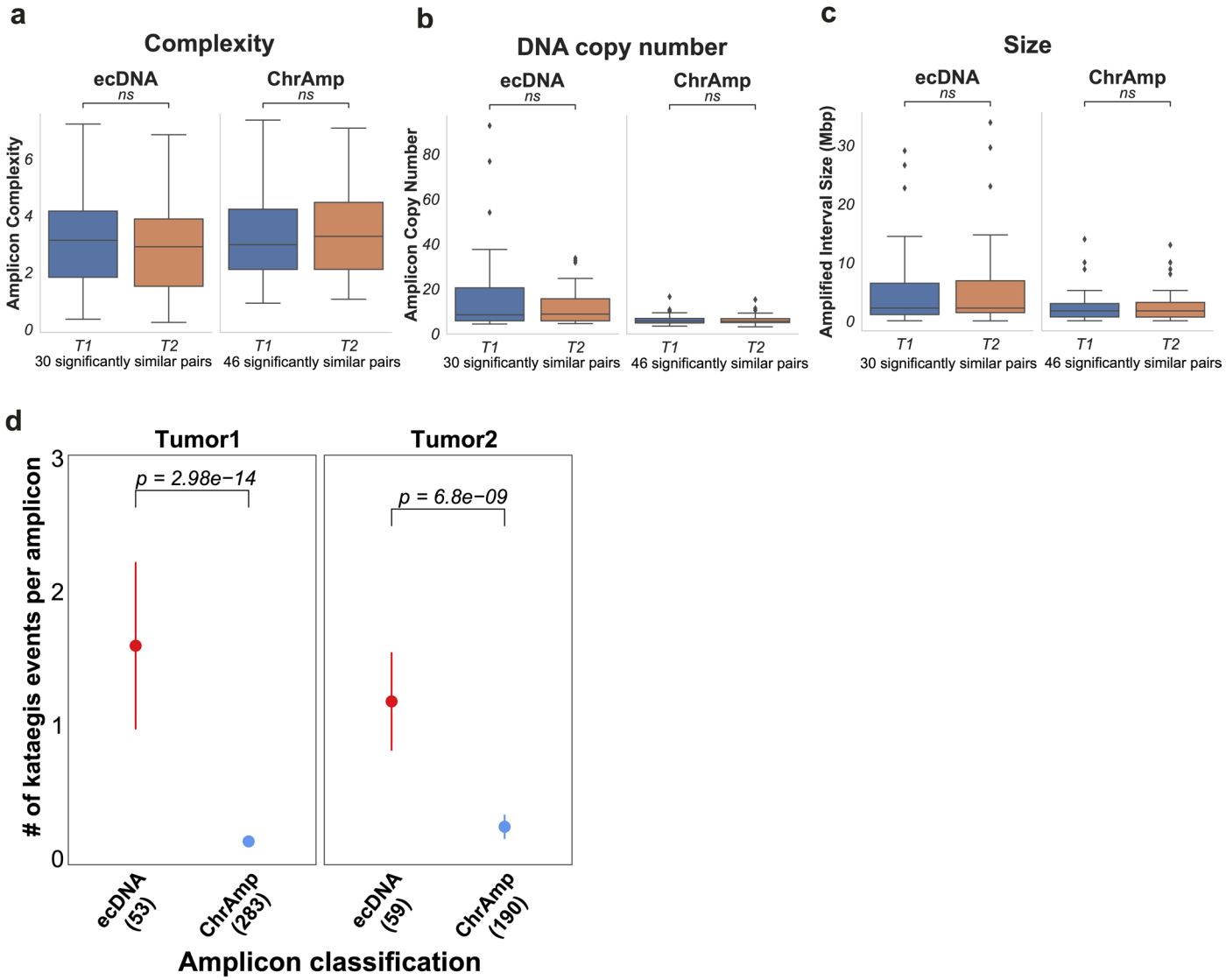
Extended Data Fig. 6 | Effects of pretreatments on distributions of sample and amplicon classifications. a, Distribution of ecDNA/ChrAmp/NoAmp tumors across the number of pretreatment a patient received. Numbers in parentheses indicate tumors with ecDNA/all tumors. P value was calculated using a two-sided Mann–Kendall trend test. **b**, Distribution of the number of distinct ecDNA amplicons pretreatment count (advanced cancers only). P value was calculated using a two-sided Mann–Kendall trend test. Points represent mean values and error bars show a 95% confidence interval. Only patients with available clinical information were included. Numbers indicate the number of patients. **c**, Distribution of ecDNA/ChrAmp/NoAmp tumors by consolidated pretreatment categories. Numbers in parentheses indicate tumors with ecDNA/all tumors. Only treatment types >50 patients are shown. P values were calculated using a two-sided binomial with the ecDNA-carrying tumor category in the untreated group as a null probability. **d**, Odds of tumors treated with targeted inhibitors to contain target oncogene on an ecDNA compared to tumors treated with targeted inhibitors lacking the amplified target, when compared to the background distribution calculated with the untreated primary tumors. **e**, EcDNA or ChrAmp amplicons by pretreatment mechanisms. Only treatments used in ≥10 patients were included. Samples were categorized solely based on whether they received chemotherapy of a specific mechanism, without considering other treatments including radiation. The points on the graph represent the mean, and the error bars indicate the standard error of the mean. The numbers shown at the bottom of the figure are sample sizes. P-values were calculated with two-sided Mann–Whitney U test. **f**, Sample classification (ecDNA, ChrAmp, NoAmp) in the advanced cohort by different pretreatment chemotherapy mechanisms. Only treatments used in ≥10 patients were included. Samples were categorized solely based on whether they received chemotherapy of a specific mechanism, without considering other treatments including radiation. As a result, the samples might have received multiple types of treatments. The p-value was calculated using a two-sided binomial test, with untreated samples serving as the reference for each chemotherapy mechanism. n.s., not significant.

https://doi.org/10.1038/s41588-024-01949-7



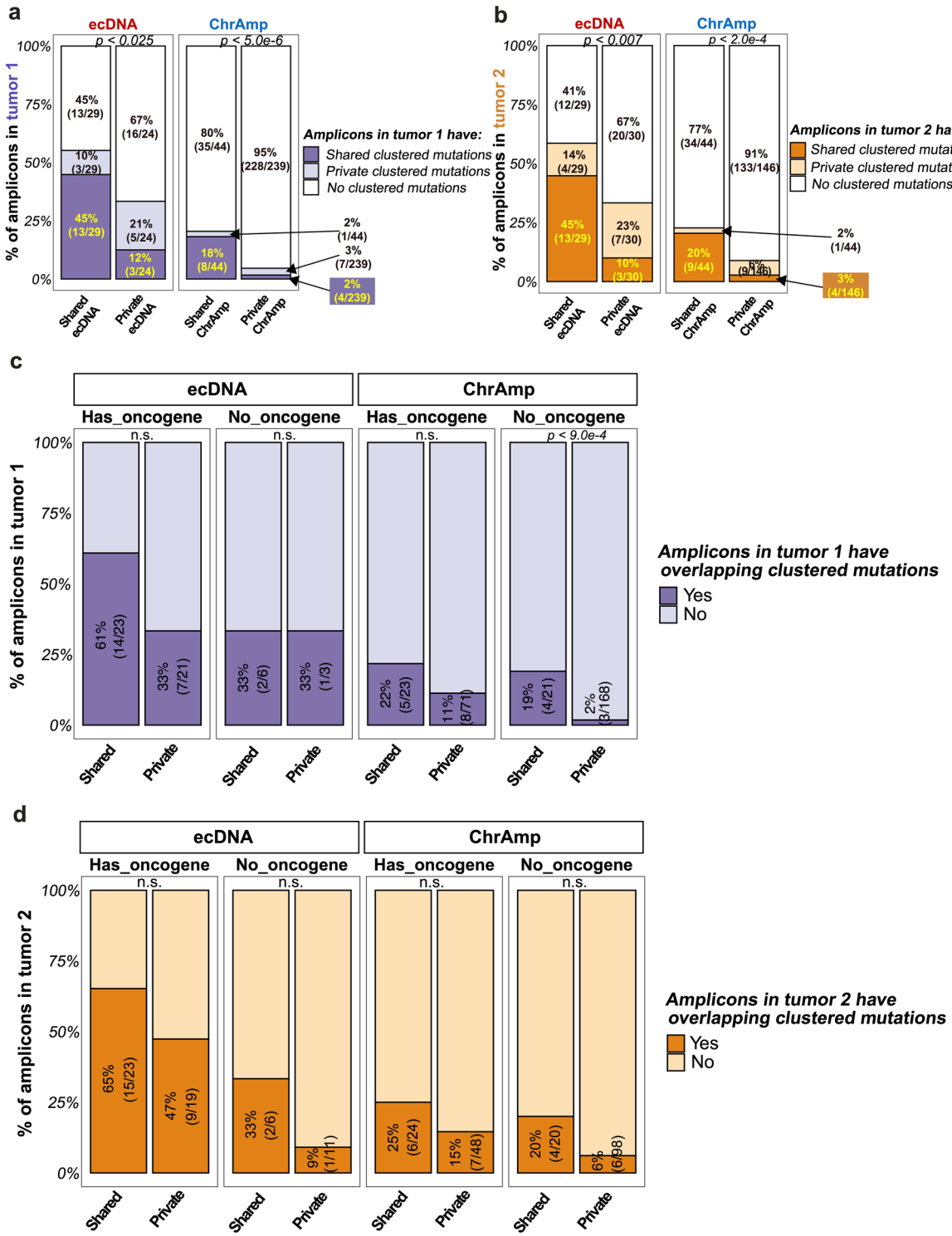
Extended Data Fig. 7 | Longitudinal analysis of sample classification. **a**, Sankey plot showing sample classification based on amplicon status, over time. Color reflects amplicon-based sample classification and numbers indicate the number of samples. **b**–**f**, Amplicon structure of five amplicons classified as ecDNA at tumor 1 (T1), and ChrAmp at tumor 2 (T2). All amplicon pairs showed a significant similarity score between T1 and T2, with T1 classified as ecDNA and T2 classified as ChrAmp. BFB, breakage fusion bridge.

https://doi.org/10.1038/s41588-024-01949-7



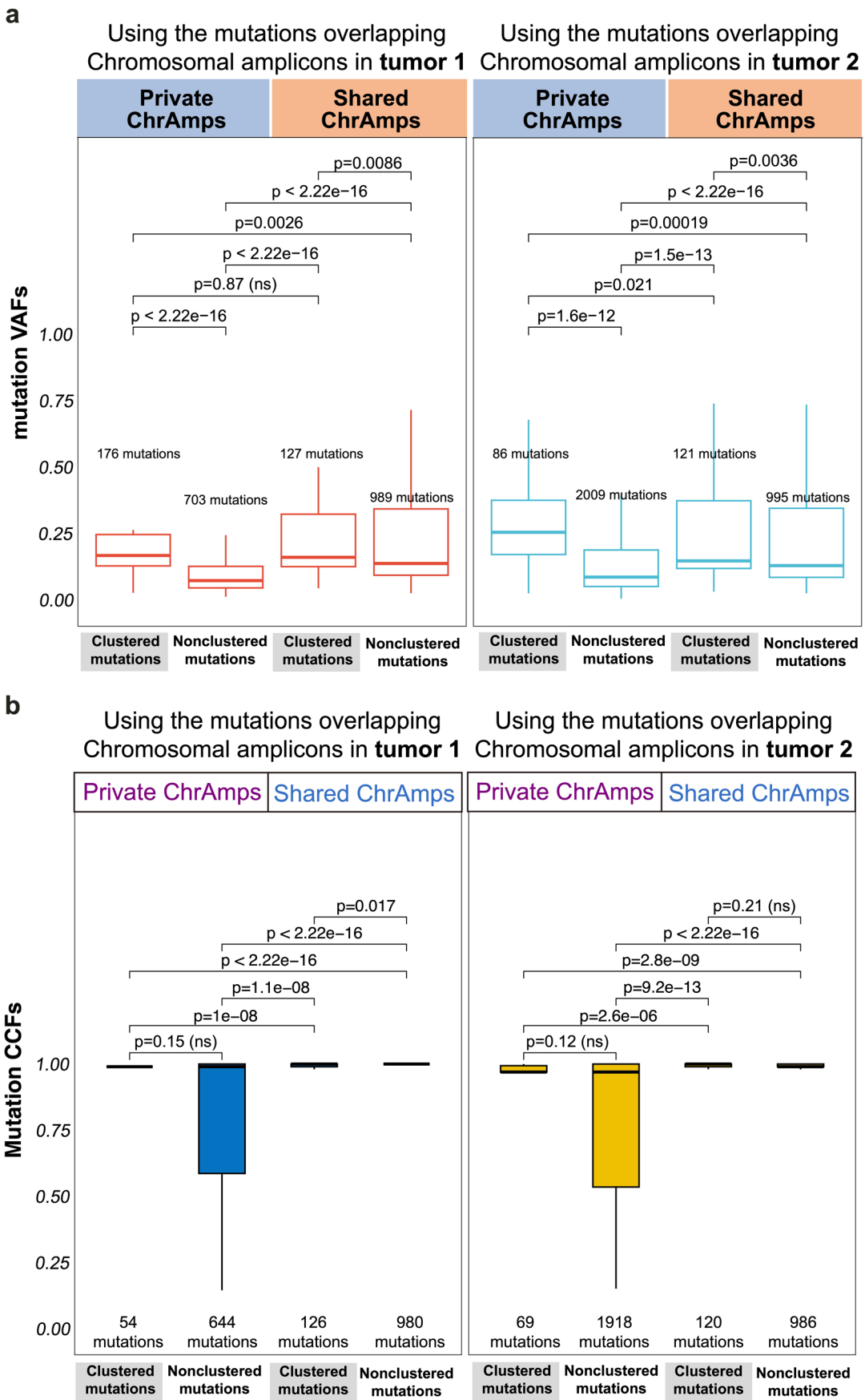
Extended Data Fig. 8 | Genomic characteristics of longitudinally retained amplicons. a–**c**, Complexity (**a**), DNA copy number (**b**) and amplicon size (**c**). P-values were computed using a two-sided Wilcoxon paired test. T1 and T2 represent a patient's first time point tumor and second-time point tumor, respectively. Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median. n.s., not significant. **d**, The number of kataegis events is significantly higher in ecDNA amplicons compared to ChrAmp amplicons, at both time points. Numbers in parentheses indicate numbers of ecDNA or ChrAmp amplicons. Error bars represent the standard error (95% confidence interval) of the mean. P values were calculated using a two-sided Mann–Whitney U test.

https://doi.org/10.1038/s41588-024-01949-7



Extended Data Fig. 9 | Additional data to longitudinal amplicon analysis. **a**, The fraction of ecDNA and ChrAmp amplicons with overlapping clustered mutations in the 1st tumor. Clustered mutations were further classified into 'shared clustered mutations' when two or more mutations in the clustered mutation event were retained in the 2nd tumor, 'private clustered mutations' when the clustered mutation event was detected in the 2nd tumor, and 'no clustered mutations' when no T1 clustered mutations were recovered in the T2 amplicon. **b**, The fraction of ecDNA and ChrAmp amplicons with overlapping clustered mutations in the 2nd tumor. Clustered mutations were further classified into 'shared clustered mutations' when two or more mutations in the clustered mutation event were retained in the 1st tumor, 'private clustered mutations' when the clustered mutation event was detected in the 1st tumor and 'no clustered mutations' when no T2 clustered mutations were recovered in the T1 amplicon. For **a** and **b**, statistical significance was assessed with chi-squared test for retained vs all others. **c**, The fraction of ecDNA and ChrAmp amplicons with overlapping clustered mutations in the 1st tumor. Numbers in parentheses indicate numbers of 1st tumor amplicon overlapping clustered mutations. **d**, The fraction of ecDNA and ChrAmp amplicons with overlapping clustered mutations in the 2nd tumor. Numbers in parentheses indicate numbers of 2nd tumor amplicon overlapping clustered mutations. P-values were computed using a chi-square test. n.s., not significant.

https://doi.org/10.1038/s41588-024-01949-7



Extended Data Fig. 10 | Additional data to variant allele fraction by mutational category. a,**b**, Comparison of (**a**) variant allele fractions and (**b**) cancer cell fractions (of different mutational categories detected on longitudinally retained (shared) or disappeared/acquired (private) ChrAmp amplicons). Boxplots represent minimum (0th percentile), maximum (100th percentile), 1st and 3rd quartiles and median with outliers excluded. P values were calculated using a two-sided Mann–Whitney U test. n.s., not significant.

nature portfolio

Corresponding author(s): Hoon Kim, Roel G.W. Verhaak

Last updated by author(s): Aug 20, 2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial 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	All data analyzed in the manuscript were obtained from publicly available resources, including The Cancer Genome Atlas, the International Cancer Genome Consortium, Hartwig Medical Foundation, and the Glioma Longitudinal Analysis Consortium. For details, please see the Data Availability Statement.
Data analysis	AmpliconSuite-pipeline (version 0.1344.2) is available at https://github.com/AmpliconSuite/AmpliconSuite-pipeline AmpliconClassifier (version 0.4.11) is a part of the AmpliconSuite-pipeline version 0.1344.2 SigProfilerClusters (version 1.0.11) is available at https://github.com/AlexandrovLab/SigProfilerClusters SigProfilerMatrixGenerator (version 1.2.15): https://github.com/AlexandrovLab/SigProfilerMatrixGenerator SigProfilerSimulator (version 1.1.4): https://github.com/AlexandrovLab/SigProfilerSimulator PyClone-VI (version 0.1.2): https://github.com/Roth-Lab/pyclone-vi R: version 4.1.2 Python: version 3.9.13

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

WGS from TCGA were accessed from the GDC (hppts://gdc.cancer.gov/) under accession ID phs000178.v11.p8 (The Cancer Genome Atlas); dbGap application # 34846. WGS data from PCAWG/ICGC were downloaded from the International Genome Consortium (ICGC) at https:/dcc.icgc.org/ (DACO application # DACO-753). The whole-genome sequencing, RNA sequencing and corresponding clinical data used in this study were made available by the HMF and were accessed under a license agreement (HMF DR-057 version 3.0). Data access can be obtained by filling out a data request form. The form and detailed application procedures can be found at https://www.hartwigmedicalfoundation.nl/applying-for-data/. Processed sequencing data from the GLASS project used in this study are available on Synapse, at https://www.synapse.org/glass. AmpliconSuite output files for TCGA are available from https://ampliconrepository.org/project/655bda68bba7c92509522479; AmpliconSuite output files for PCAWG are found at https://ampliconrepository.org/project/655c060abba7c925095555da.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex and gender were not considered when designing this study.
Reporting on race, ethnicity, or other socially relevant groupings	race, ethnicity or other socially relevant groupings were not considered when designing this study.
Population characteristics	Detailed population characteristics for HMF, TCGA-ICGC, GLASS have been described in their marker papers: TCGA-ICGC: The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Nature 2020; PMID 32025007. HMF cohort: Priestley et al., Nature, 2019 ; PMID 31645765. GLASS: Barthel et al., Nature 2019; PMID 31748746. In brief, the HMF cohort consists of 1858 (50.4%) males and 1832 (49.6%) females, when excluding 480 patients with unknown sex), with mean age of 61.4 years ranging from 17-89. 2333 (66%) of the patients were pretreated with either systemic-therapy or ratio-therapy. The TCGA-ICGC cohort consists of 1985 (54.0%) males and 1688 (46.0%) females with mean age of 57.0 years ranging from 1-90.1. The patients are in principle primary cancers that, in most cases, were not pretreated before biopsy. The GLASS cohort includes 100 patients of two time-point tumor samples that were mostly pretreated before the second time-point biopsy. It consists of 59 (59%) males and 41 (41%) females) with mean age of 47.8yrs ranging from 21~71 years old.
Recruitment	Recruitment was performed outside the scoped of the current study and as detailed in the original publications.
Ethics oversight	Hartwig Medical Foundation, TCGA-ICGC and GLASS; controlled access to Hartwig, TCGA-ICGC was deemed exempt of IRB oversight by the IRB at Yale University. Not yet published GLASS datasets included longitudinal sample pairs of glioma and glioblastoma tumors, collected from the Centre Hospitalier de Luxembourg (CHL, Neurosurgical Department) from patients who had given their informed consent. The study received official approval from the National Committee for Ethics in Research (CNER) Luxembourg, under the protocol number 201201/06. Additional longitudinal sample pairs of glioma and glioblastoma tumors were collected from the Department of Neurosurgery, Seoul National University Hospital. It was approved by the Institutional Review Board of Seoul National University Hospital (approval number H-2004-049-1116), and all patients provided signed informed consent accordingly.

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

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 collected whole genome sequencing for all possible samples from TCGA-ICGC, HMF, and GLASS cohorts. The TCGA cohort included 1940 patients out of which 858 patients were also present in the ICGC which consisted of 2785 patients. The HMF cohort included 4172 patients.

	After filtering patients (see Extended Data Figure 1.C), we selected 7860 patients (3690 TCGA-ICGC, 4170 HMF) each of which consisted of a single tumor and its patient-matching normal sample. The results in our manuscript are derived from statistical tests that take sample sizes into account when determining significance, negating the need for an a priori sample size determination. No statistical methods were used to predetermine sample size.
Data exclusions	All available whole genome sequencing data from HMF, TCGA-ICGC and GLASS were analyzed. No samples were excluded a priori.
Replication	Independent analyses resulted in repeated results, verifying the reproducibility of our findings.
Randomization	Samples from TCGA-ICGC, HMF, and GLASS cohorts were independently collected by their data collection centers. When comparing primary vs advanced tumors, we ensured that both groups had a sufficient number of tumors per tumor type in order to account for potential disproportinate samples distribution over different tumor types.
Blinding	The samples we analyzed were deidentified by HMF, TCGA-ICGC and GLASS

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*