Extrachromosomal DNA (ecDNA) amplification promotes intratumoral genetic heterogeneity and accelerated tumor evolution; however, its frequency and clinical impact are unclear. Using computational analysis of whole-genome sequencing data from 3,212 cancer patients, we show that ecDNA amplification frequently occurs in most cancer types but not in blood or normal tissue. Oncogenes were highly enriched on amplified ecDNA, and the most common recurrent oncogene amplifications arose on ecDNA. EcDNA amplifications resulted in higher levels of oncogene transcription compared to copy number-matched linear DNA, coupled with enhanced chromatin accessibility, and more frequently resulted in transcript fusions. Patients whose cancers carried ecDNA had significantly shorter survival, even when controlled for tissue type, than patients whose cancers were not driven by ecDNA-based oncogene amplification. The results presented here demonstrate that ecDNA-based oncogene amplification is common in cancer, is different from chromosomal amplification and drives poor outcome for patients across many cancer types.

Somatic gain-of-function alterations in driver oncogenes play a central role in cancer development. Oncogene amplification is the most common gain-of-function alteration in cancer, enabling tumor cells to circumvent the checks and balances that are in place during homeostasis to drive tumor growth. EcDNA-based amplification has long been recognized as a way for cells to increase the copy number (CN) of specific genes, but their frequency is unclear. EcDNA amplification may enable tumors to reach high CN of growth-promoting genes, while maintaining intratumoral genetic heterogeneity through its nonchromosomal mechanism of inheritance. To date, low-throughput cytogenetic methods have been used to infer extrachromosomal status of DNA amplifications.

Consequently, the frequency, distribution and clinical impact of ecDNA-based amplification have not been comprehensively assessed. More recently, computational analyses of whole-genome sequencing (WGS) data and new circular DNA library enrichment approaches have suggested a relatively high frequency of ecDNA, particularly in tumors of the central nervous system. In this study, we set out to perform a global survey of the frequency of ecDNA-based oncogene amplifications, while investigating their contents and determining their clinical context.

We leveraged three characteristic properties of ecDNA to enable our computational analysis: (1) ecDNA are circular; (2) they are highly amplified; and (3) they lack a centromere. These properties provide a basis for the AmpliconArchitect tool, which enables the detection and characterization of ecDNA from WGS data. We applied AmpliconArchitect to WGS data from tumor tissue, matched normal tissue and blood from The Cancer Genome Atlas (TCGA) and tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) (ref. 1) to quantify and characterize the architecture of regions that are larger than 10 kilobases (kb) and have more than 4 copies (CN > 4) above median sample ploidy (Supplementary Table 1). Amplicons were classified as: (1) ‘circular’ (Fig. 1a), representing amplicons residing extrachromosomally; (2) ‘BFB’ if they bore a signature of having been created by a breakage-fusion-bridge (BFB) mechanism; (3) as ‘heavily rearranged’; for noncircular amplicons containing pieces of DNA (DNA segments) from different chromosomes or regions that were very far apart on chromosomes (>1 Mb); or (4) ‘linear’ for linear amplifications. Amplicon status provided the basis for classification of tumor samples. Samples lacking amplifications were labelled ‘no focal somatic CN amplification detected’.

To evaluate the accuracy of the AmpliconArchitect predictions, we analyzed WGS data from a panel of 44 cancer cell lines and
Fig. 1 | Frequency of circular amplification across tumor and nontumor tissues. a, Schematic representation of the four classification categories. All DNA regions with a CN of 4 or greater than ploidy and comprising at least 10 kb were classified using a hierarchical scheme based on the AmpliconArchitect amplicon reconstruction and the types of discordant breakpoint edges in the region. The four categories are defined as follows: linear amplicon, that is, an amplicon that contains amplified segments with either no discordant edges or with edges suggesting deletions smaller than 1 Mb; heavily rearranged amplicon, that is, an amplicon that contains amplified segments connected by discordant breakpoint edges suggesting higher-order rearrangements beyond small deletions such as inversions, interchromosomal edges or deletions >1 Mbp; BFB amplicon, that is, an amplicon having a proportion of foldback reads in excess of 25% and which may have signatures of heavily rearranged or circular amplification; circular amplicon, that is, an amplicon that contains one or more genomic segments forming a cyclic path of at least 10 kbp and 4+ copies.

b, Left: comparison of WGS-derived circular DNA amplicon and CIRCLE-seq-derived segments. Right: circular amplicons detected from WGS with AmpliconArchitect were validated with CIRCLE-seq. c, Distribution of circular, BFB, heavily rearranged, linear and no focal somatic CN amplification detected amplicon categories by tumor and normal tissue across 3,731 tumor and nonneoplastic sample-derived whole genomes from the TCGA and 1,291 whole genomes from the PCAWG.
examined tumor cells in metaphase. We used 35 unique FISH probes in combination with matched centromeric probes (81 distinct ‘cell line-probe’ combinations) to determine the intranuclear location of amplicons (Supplementary Table 2). Following automated analysis >1,600 images, we observed that 85% of amplicons characterized as ‘circular’ by WGS profile demonstrated an extra-chromosomal fluorescence signal, representing the positive predictive value. Of the amplicons corresponding to extrachromosomally located FISH probes, 83% were classified as circular, representing the sensitivity (Extended Data Fig. 1a). Circular amplicons had a median count of 16.6 ecDNA per cell, compared to 0.1 ecDNA per cell for other amplicon classes combined (collectively referred to as ‘noncircular’). In the subset of amplicons classified as circular, 6 of 34 contained co-occurring extrachromosomal and chromosomal signals, suggesting that ecDNA may coexist with ecDNA that has reintegrated into the genome14,15.

To validate our amplicon classification in patient tumors, we classified amplicons detected in the WGS data from 15 neuroblastomas and compared these to circularization for in vitro reporting of cleavage effects by sequencing (CIRCLE-seq) results, a sequencing library enrichment approach optimized for circular DNA detection16-20. We observed a very high concordance between WGS and CIRCLE-seq approaches in distinguishing circular from linear DNA amplicons (Fig. 1b and Extended Data Fig. 1b–d). AmpliconArchitect classified 4 of 65 amplicons as circular; all 4 were validated by CIRCLE-seq. No CIRCLE-seq reads were detected in 60 of the 61 remaining non-circular amplicons. One of the amplicons detected by CIRCLE-seq was classified as noncircular by AmpliconArchitect. Together with the cell line-based validation, these results suggest that our classification of WGS-derived amplifications is sensitive and has a high positive predictive value.

Having observed that we can specifically detect extrachromosomal amplifications, we applied AmpliconArchitect classification to the WGS datasets from 3,212 tumor and 1,810 nonneoplastic samples, comprising 3,212 patients (Supplementary Table 3). We found that 460 (14.3%) tumor samples carried 1 or more circular amplicons, demonstrating that ecDNA-based amplification is a common event in cancer (Fig. 1c). In contrast, circular amplifications were nearly undetectable in matched whole-blood or normal tissue samples (Fig. 1c). Of note, our analysis does not reflect the presence of cell-free DNA in blood or of small (<1 kb), circular, nonamplified DNAs, which are common both in nonneoplastic and tumor tissues20-22. EcDNA-based circular amplicons were found in 25 of 29 cancer types analyzed, including at high frequency in aggressive histological cancers such as glioblastoma, sarcoma and esophageal carcinoma. The distribution of circular amplicon frequencies is consistent with earlier results on cancer models2-3 and showed that ecDNA amplifications are a defining feature of multiple cancer subtypes but not normal cells.

The chromosomal distribution of the 579 circular amplicons was highly nonrandom (Fig. 2a), more so when compared to chromosomal regions from noncircular classes. We found that 38% of the 24 most recurrent amplified oncogenes were most frequently present on circular amplicons, with frequencies ranging from 11% of samples for Pax8 to 62% for Cdk4 (Fig. 2b and Extended Data Fig. 2a). The result carried over to a larger list of 1,804 oncogenes that were amplified in at least 5 samples, with 21.8% of those oncogenes having a plurality for being amplified on circular structures (Extended Data Fig. 2b). For highly amplified oncogenes (that is, CN > 8), the proportion further increased to 53.5%. Oncogenes amplified on circular amplicons achieved higher CNs than the same oncogenes amplified on noncircular structures (Extended Data Fig. 2c). We further observed that the association between ecDNA structures and oncogene amplification did not extend to breakpoints. For 24 frequently amplified oncogenes, the frequency of observing a specific number of breakpoints in a unit interval decayed exponentially, which is consistent with mostly random occurrence around the oncogene (Fig. 2c, Extended Data Fig. 2d and Extended Data Fig. 2e).

These results suggest that ecDNA is formed through a random process, where selection for higher copies of growth-promoting oncogenes leads to rapid oncogene amplification during cancer development, retaining intratumoral genetic heterogeneity due to its uneven inheritance24.

We compared amplicon classes for different types of genomic instability. Circular and noncircular amplifications showed similar likelihood of occurring in samples with chromosome arm-level aneuploidy (Extended Data Fig. 3a) and whole-genome duplication (Extended Data Fig. 3b), which might arise as a result of chromosome missegregation25 or other mitotic errors26. Smaller, focal genomic gains and losses result from different mutagenic processes than aneuploidy events. We observed an increase in the genome-wide number of DNA segments in samples marked by circular amplicons compared to other categories (Fig. 2d). The frequency of CN losses was comparable between circular and noncircular amplicon class samples (Extended Data Fig. 3c), but genomic segment gains were more frequently detected in samples with circular amplification compared to noncircular amplicon class samples (Wilcoxon rank-sum test, \( P < 0.03 \) for BFB, \( P < 0.03 \) for heavily rearranged, \( P < 1 \times 10^{-10} \) for linear and \( P < 1 \times 10^{-11} \) for no focal somatic CN amplification detected; Extended Data Fig. 3d). Most circular amplicon breakpoints showed no or minimal sequence homology (<5 bp), implicating nonhomologous end joining in ecDNA-associated breakpoint repair. In contrast, noncircular amplicon breakpoints showed significantly more microhomologies (Extended Data Fig. 3e, \( P < 1 \times 10^{-14} \); two-sided Fisher’s exact test). Nonhomologous end joining has been associated with localized breakpoint clustering or chromothripsis27.

Somatic structural aberrations, such as chromothripsis, do not cause amplification but may create circular structures that can be subsequently amplified. We detected signatures of chromothripsis in 36% of circular amplicons (Extended Data Fig. 3f) and half of circular amplicon cases (Extended Data Fig. 3g). The prevalence of chromothripsis was higher among the circular class than other classes (chi-squared \( P = 2.2 \times 10^{-10} \)). This result confirms recent observations that chromothripsis can result in BFB and ecDNA formation28-30 and identifies chromothripsis as an initiating event for some ecDNAs. In contrast, genome-wide tandem duplications31 were not associated with ecDNA (chi-squared \( P = 0.1 \); Extended Data Fig. 3h).

We sought to examine the transcriptional consequences of circular ecDNA amplification. As expected, we observed a highly significant correlation between DNA CN and oncogene expression level in all amplicon categories. However, when normalized for DNA CN, oncogenes on circular amplicons showed significantly higher expression than noncircular amplicon oncogenes (1.2x higher compared to noncircular amplifications, \( P < 0.0007 \); Tukey’s range test; Fig. 3a and Extended Data Fig. 4). The CN-independent increase in transcriptional activity may be in part the result of enhancer hijacking events and enhanced chromatin accessibility on ecDNA elements32,33. To compare the epigenetic mechanisms governing gene expression between circular amplifications and noncircular regions, we analyzed the overlapping assay for transposase-accessible chromatin using sequencing (ATAC-seq) profiles available for 36 samples34. Following DNA CN level correction, the chromatin of circular and BFB amplicons was significantly more accessible compared to heavily rearranged and linear amplicons (1.2x higher median ATAC-seq signal fold change; Wilcoxon rank-sum test; \( P < 1 \times 10^{-14} \); Fig. 3b), consistent with recent findings that increased accessibility plays a role in the dysregulation of ecDNA oncogenes35,36,37. Finally, the frequency of amplicon-derived transcript fusions was increased fivefold in circular compared to noncircular amplifications (Fig. 3c; binomial test, \( P < 1 \times 10^{-14} \)). We observed a convergence of DNA
Fig. 2 | Oncogene content and structural component of circular amplification. a, Genome-wide distribution of amplification peaks by amplicon class. Amplifications were counted per 1 Mb bin and are shown as a fraction of the total number of samples per amplicon class. b, Classification of amplification status by gene. The 24 most frequently amplified oncogenes are shown. c, Breakpoint locations (right) and distribution of breakpoints (left) across all circular samples with amplified CCND1, EGFR and MYC. Breakpoints were identified in each sample containing the amplified oncogene region. The total number of breakpoints across this region in 1-kb binned windows (right) are shown. The distribution of the number of breakpoints in each bin closely follows a Poisson distribution (left), suggesting that the breakpoints are mostly randomly distributed across the region. d, The number of genome-wide DNA segments within a sample was compared between circular, BFB, heavily rearranged, linear and no focal somatic CN amplification detected classes. Circular samples contained statistically significantly more DNA segments than noncircular samples (P = 0.0046, 7.2 x 10^{-6}, 2.4 x 10^{-15} and 9.4 x 10^{-125}, respectively; two-sided Wilcoxon rank-sum test).
CN, RNA expression and chromatin accessibility around circular amplicon structures (Fig. 3d).

To determine whether cancers that have ecDNA amplification were associated with aggressive biological features, we examined the impact of circular amplification on lymph node status. Gene amplification, whether ecDNA-based, by BFB formation or heavily rearranged, was associated with significantly more lymph node spread at initial diagnosis (chi-squared test, $P < 1 \times 10^{-15}$) (Extended Data Fig. 5). To further examine the association of ecDNA with biological features of aggressiveness, we used gene expression signatures of increased tumor cell proliferation and reduced immune cell infiltration$^{14}$. The cellular proliferation scores of the circular
and BFB class samples were significantly higher ($P < 1 \times 10^{-15}$; Wilcoxon rank-sum test; Extended Data Fig. 6a) compared to the heavily rearranged, linear and no focal somatic CN amplification detected categories. Accordingly, the linear and no focal somatic CN amplification detected groups showed higher immune infiltration scores compared to the circular, BFB and heavily rearranged samples (Extended Data Fig. 6b; $P < 1 \times 10^{-4}$; Wilcoxon rank-sum test). Combined, these scores suggest that tumors carrying a circular amplicon behave aggressively.

Most importantly, patients whose tumors contained at least one circular amplicon had significantly worse outcomes compared to patients whose tumors were classified as noncircular. The $P$ value derived from comparing the survival curves was based on a log-rank test. B. Multivariate Cox proportional-hazards model, incorporating disease and patient cohorts as the parameters, showing that circular amplification resulted in significantly higher hazard ratios. The error bars represent the 95% confidence intervals of the hazard ratios.

A map of the cancer genome that respects only the direct changes to its ‘genetic code’ and not also genome topology and three-dimensional organization, will be necessarily incomplete. Three-dimensional genome topology plays a critical role in determining how that genome functions, or malfunctions, as occurs in cancer. Detection and classification of circular ecDNA creates a more accurate map of the cancer genome. The data presented in this study demonstrate that circular ecDNA play a critical role in cancer, providing a mechanism for achieving and maintaining high copy oncogene amplification and diversity while driving enhanced chromatin accessibility and elevating oncogene transcription. This mechanism of amplification is operant in a large fraction of human cancers and negatively affects patient outcomes, independent of cancer lineage. Our results present the landscape of ecDNA across cancer. Given cancer’s heterogeneity, it is certain that diversity in ecDNA structure and behavior exists between different cancer types. Future studies, such as deep dives into patterns of complex structural variation...
across cancer\textsuperscript{24,51,52}, will aid improved understanding of the mechanisms that create genomic rearrangements including ecDNA. The potential to leverage the presence of ecDNAs in human cancers for diagnostic or therapeutic purposes provides a link between cancer genomics and broad utility for patient populations.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0678-2.

Received: 17 June 2020; Accepted: 15 July 2020; Published online: 17 August 2020

References
1. de Carvalho, A. C. et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. \textit{Nat. Genet.} \textbf{50}, 708–717 (2018).
2. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. \textit{Nature} \textbf{543}, 122–125 (2017).
3. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. \textit{Nat. Rev. Cancer} \textbf{19}, 283–288 (2019).
4. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. \textit{Nat. Genet.} \textbf{49}, 65–74 (2017).
5. Zack, T. T. et al. Pan-cancer patterns of somatic copy number alteration. \textit{Nat. Genet.} \textbf{45}, 1134–1140 (2013).
6. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. \textit{Nature} \textbf{463}, 899–905 (2010).
7. Alt, F. W., Kellemes, R. E., Bertino, J. R. & Schimke, R. T. Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. \textit{J. Biol. Chem.} \textbf{253}, 1357–1370 (1978).
8. Kohl, N. E. et al. Transposition and amplification of oncogene-related sequences in human neuroblastomas. \textit{Cell} \textbf{35}, 359–367 (1983).
9. Nathanson, D. A. et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. \textit{Science} \textbf{343}, 72–76 (2014).
10. Zheng, S. et al. A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. \textit{Genes Dev.} \textbf{27}, 1462–1472 (2013).
11. Trask, B. J. Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. \textit{Trends Genet.} \textbf{7}, 149–154 (1991).
12. Deshpande, V. et al. Exploring the landscape of focal amplifications in cancer using AmpliConArchitect. \textit{Nat. Commun.} \textbf{10}, 392 (2019).
13. Xu, K. et al. Structure and evolution of double minutes in diagnosis and relapse brain tumors. \textit{Acta Neuropathol.} \textbf{137}, 123–137 (2019).
14. Koche, R. P. et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. \textit{Nat. Genet.} \textbf{52}, 29–34 (2020).
15. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. \textit{Nature} \textbf{578}, 82–93 (2020).
16. Zakov, S., Kinsella, M. & Bafna, V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. \textit{Proc. Natl. Acad. Sci. USA} \textbf{110}, 5546–5551 (2013).
17. Rajkumar, U. et al. EcSeg: semantic segmentation of metaphase images containing extrachromosomal DNA. \textit{iScience} \textbf{21}, 428–435 (2019).
18. Storlazzi, C. T. et al. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. \textit{Genome Res.} \textbf{20}, 1198–1206 (2010).
19. Møller, H. D. et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. \textit{Nat. Commun.} \textbf{9}, 1069 (2018).
20. Kumar, P. et al. Normal and cancerous tissues release extrachromosomal circular DNA (ecDNA) into the circulation. \textit{Mol. Cancer Res.} \textbf{15}, 1197–1205 (2017).
21. Shibata, Y. et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. \textit{Science} \textbf{336}, 82–86 (2012).
22. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. \textit{Annu. Rev. Cell Dev. Biol.} \textbf{27}, 585–610 (2011).
23. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. \textit{Nat. Genet.} \textbf{50}, 1189–1195 (2018).
24. Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. \textit{Nat. Genet.} \textbf{52}, 331–341 (2020).
25. Li, P. et al. Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. \textit{Nat. Genet.} \textbf{51}, 705–715 (2019).
26. Zhang, C.-Z. et al. Chromothripsis from DNA damage in micrornuclei. \textit{Nature} \textbf{522}, 179–184 (2015).
27. Umbreit, N. T. et al. Mechanisms generating cancer genome complexity from a single cell division error. \textit{Science} \textbf{368}, eaab0712 (2020).
28. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. \textit{Cell} \textbf{179}, 1330–1341.e13 (2019).
29. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. \textit{Science} \textbf{362}, eaax1898 (2018).
30. Helmsauer, K. et al. Enhancer hijacking determines intra- and extrachromosomal circular MYCN amplicon architecture in neuroblastoma. Preprint at bioRxiv https://doi.org/10.1101/2019.12.20.875807 (2019).
31. Wu, S. et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. \textit{Nature} \textbf{575}, 699–703 (2019).
32. Conneely, K. M. & Dean, M. A. Genes and mechanisms of breast cancer: an update. \textit{Nat. Rev. Cancer} \textbf{3}, 203–214 (2003).
33. Helmsauer, K. et al. Enhancer hijacking determines intra- and extrachromosomal circular MYCN amplicon architecture in neuroblastoma. Preprint at bioRxiv https://doi.org/10.1101/2019.12.20.875807 (2019).
34. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. \textit{Science} \textbf{355}, eaav8399 (2017).
35. Hadi, K. et al. Novel patterns of complex structural variation revealed across thousands of cancer genome graphs. Preprint at bioRxiv https://doi.org/10.1101/836296 (2019).
36. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. \textit{Nature} \textbf{575}, 210–216 (2019).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. © The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
AmpliconArchitect takes as input an aligned WGS sequences and seed intervals for a candidate amplicon region. AmpliconArchitect then searches for other regions that belong to the amplicon by exploring the seed intervals and extends beyond the intervals if it encounters CN changes or discordant edges that support a breakpoint. The collection of intervals and breakpoints are combined to form a breakpoint graph with nodes representing segments and edges representing rearrangements. This breakpoint graph can be further decomposed into simple and complex cycles to identify any circular paths within the amplicon structure, which is indicative of an ecDNA origin. AmpliconArchitect masks out regions that are highly repetitive, including the alpha satellites seen in centromeric and pericentromeric regions. Therefore, they are not part of the amplicon structure. While AmpliconArchitect sees amplicons that may reside on chromosomes, the predicted circular structures will not include centromeres. The detected amplicons were annotated with the Ensembl (release 75) gene database (GRCh37).

BFB. BFB status was determined by evaluating the AmpliconArchitect output. We examined the AmpliconArchitect graph files to identify amplicons with a proportion of foldback breakpoint edges exceeding 0.25 and having at least 25 sequencing reads supporting all edges in the graph. Foldback breakpoint edges were defined as AmpliconArchitect breakpoint edges whose constituent sequencing reads had forward and reverse mates in the read pair with the same orientation or opposite to the expected (+/+ or −/− or +/−), and for which the edge spanned ≤25 kbp in the reference genome. Amplicons meeting these criteria were classified as BFB. We note that our approach is probably only identifying linear BFB amplicons, not circular BFB structures.

Amplicon and sample classification. As a prerequisite, amplicons must contain >10 kb of genomic segments amplified to at least 4 copies above median ploidy to be considered a valid amplicon. We then used the AmpliconArchitect-derived breakpoint graph to classify amplicons into four categories: (1) circular amplification; (2) BFB amplification; (3) heavily rearranged amplification; and (4) linear amplification (Fig. 1a). Amplicons were denoted as circular amplification if the segments formed a cycle in the graph of a total size of at least 10 kb and had at least a copy count of 4. Amplicons were denoted as BFB if they met the criteria for a BFB amplicon. Since cyclic structures can arise in the breakpoint graph of a linear BFB due to repetitive self-inversion, BFB amplicons that also contained circular amplicon signatures were classified as BFB. Noncircular amplicons were denoted as heavily rearranged if they contained amplified segments connected by discordant breakpoint edges, suggesting higher-order rearrangements beyond small deletions, such as inversions, interchromosomal edges or deletions >1 megabase pair (MbP). (Fig. 1a). Noncircular amplicons were denoted as linear if they contained amplified segments with either no discordant edges or with edges suggesting deletions <1 Mb. While an amplicon may fit the requirements for several categories (that is, a circular amplicon may also comprise heavily rearranged amplifications), priority was given to the BFB amplification category, followed by circular, heavily rearranged and then linear. Samples were classified based on which amplicons were present in the sample, giving precedence to the presence of amplicons with the highest priority, with the exception that a non-BFB circular amplicon took precedence over any of the same sample containing a BFB. For example, samples with both circular and heavily rearranged amplification would be classified as circular. Samples without any amplicons were classified as ‘no focal somatic CN amplification detected’.

Cell line validation. We ran AmpliconArchitect on the WGS data from 44 cell line models and FISH in parallel, including those described previously\(^1\). For AmpliconArchitect, the seed interval for each cell line included the probe region. For each FISH probe, we reported whether it landed in an amplicon (inferred from AmpliconArchitect) and if so, what was the amplicon classification. The distribution of the average ecDNA per cell was computed as the average number of FISH probes that colocalized on ecDNA across all the images for that particular cell line + FISH probe combination (Extended Data Fig. 1a). A Wilcoxon rank-sum test was used to detect significant differences in average ecDNA counts per cell across the amplicon classes.

We used ecSeg\(^7\) to validate the ecDNA counts and oncogene amplification on ecDNA from the cell line image data. ecSeg takes as input DAPI + FISH-stained metaphase images and uses the DAPI signal to calculate the DNA signatures as nuclear, chromosomal or extrachromosomal. It then localizes the red and green FISH signals present in the image to identify whether they are present on chromosomal or ecDNA segments. An oncogene is considered to be located on ecDNA only if the FISH signal for that oncogene is colocalized with an ecSeg classified ecDNA segment. For each image, ecSeg reports the number of times an oncogene is found on ecDNA. We report the average of these counts for each combination of cell line and FISH probe. A cell line was considered to be ecDNA-positive by FISH if it contained an average of at least 0.5 ecDNA + FISH colocalized signal per cell. All images analyzed can be obtained from figshare at https://figshare.com/s/6c3e2edc1ab2999bb2af0 and https://figshare.com/s/ab6a214738a43833391.

TCGA processing. We processed TCGA WGS BAM files through the Institute for Systems Biology 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 found at the Genomic Data Commons (https://portal.gdc.cancer.gov/legacy-archive/search/1) and the PancanAtlas publications (https://gdac.broadinstitute.org/about-data/publications/pancanatlas). We used genome-wide SNP6 CN segments with a CN log ratio ≥ 1 as seed interval(s) of interest for the input to AmpliconArchitect\(^7\). Default parameters and reference files were used for all other settings. Details on how to run AmpliconArchitect have been described in the corresponding manuscript\(^7\) and its source code repository. We ran AmpliconArchitect on tumor and normal WGS samples from 1,921 patients (3,731 BAM files). Samples were classified as BFB on the amplicon with the highest precedence present in the sample or classified as ‘no focal somatic CN amplification detected’ if no amplicons were present in the sample.

PCAWG processing. PCAWG WGS BAM files are available on the Amazon Web Services Cloud. DNA CN profile structural variant and fragments per kilobase of transcript per million mapped reads (FPKM) data were obtained from the International Cancer Genomics Consortium (ICGC) data portal at https://dcc.icgc.org/releases/PCAWG. AmpliconArchitect used CNs equal to or higher than 4 as seed interval(s) of interest. We ran AmpliconArchitect on tumor WGS samples from 1,291 patients and the results were processed in the same way as the results from TCGA.

Oncogene analysis. We examined the enrichment of the 24 recurrent oncogenes known to be activated by amplification by counting the total number of times the amplicon classes overlapped the 24 recurrent oncogenes. We then simulated 10,000 random classes by sampling random regions of the same size. The 24 recurrent oncogenes and computed an empirical expected distribution of the times these random regions overlapped with the amplicon classes. We report the z-score between the empirical distribution and observed value for the amplicon classes. We also report the average copy count estimated from AmpliconArchitect. For each of these oncogenes on an amplicon structure, we modeled a proportion of the breakpoint detected within a 1-Mb region flanking the oncogene using the breakpoint graph to infer breakthroughs. We partitioned the region into 1,000 bp windows and counted the total number of breakpoints that landed in each window; we displayed a histogram of these counts. We modeled the histograms using an exponential distribution and showed that under the assumption that the breakpoints were distributed randomly, the histograms closely followed the exponential distribution. We used allOnco (http://www.bushmanlab.org/links/genelists), a set of 2,579 cancer genes generated from curated collections of cancer genes from many different publications. We identified all amplicons that overlapped with the oncogenes and report the proportion of amplified oncogenes that are circular.

Breakpoint detection. For each of these oncogenes on an amplicon structure, we report the position of breakpoints detected within a 1-Mb region flanking the oncogene using the breakpoint graph to infer breakpoints. We partitioned the region into 1,000 bp windows and counted the total number of breakpoints that landed in each window; we displayed a histogram of these counts. For each oncogene, we modeled a proportion of breakpoints detected within a 1-Mb region flanking the oncogene using the breakpoint graph to infer breakpoints. We partitioned the region into 1,000 bp windows and counted the total number of breakpoints that landed in each window; we displayed a histogram of these counts. We modeled the histograms using an exponential distribution and showed that under the assumption that the breakpoints were distributed randomly, the histograms closely followed the exponential distribution.

Genomic instability analyses. We computed total CN gains/losses as the number of WGS-inferred CN segments with CN > 2 or CN < 2. A Wilcoxon rank-sum test was used to test for a significant difference between the two distributions. We used data from Taylor et al.\(^7\) on genome doubling status and chromosomal arm amplification and loss for each sample. A Wilcoxon rank-sum test was used to test the significance between the distribution of gains and losses; a chi-squared test was used to test the significance between the distribution of whole-genome doublings.

Transcript fusions were downloaded from the TCGA fusion database (https://tumorfusions.org/)\(^8\), \(^9\), derived using PRADA\(^7\), to identify fusion events that occurred on an amplicon. For each fusion in the database, we considered it valid if both ends of the fusion breakpoint junction occurred on the same amplicon. In total, 710 amplified fusions were detected. We computed the average fusion frequency per 10 MB as the total number of fusions that landed within an amplicon class divided by the sum of all the base pairs of the amplicon class multiplied by 1 \times 10^9. To test whether circular amplicons were enriched fusion events, we computed the \(P\) value of observing at least the number of fusion events occurring on circular amplicons under a binomial distribution where the probability \(P\) was estimated using the total number of fusion events on the amplified noncircular event divided by the total base pairs of the amplified noncircular event and divided by the number of trials \(n\) as the total base pairs of the circular amplicons.
RNA sequencing (RNA-seq) and ATAC-seq analyses. Of the 3,212 tumor samples, 2,148 had RNA-seq data in the format of FPKM upper quartile expression data. For each gene within each disease cohort, we computed a baseline FPKM upper quartile as the average FPKM upper quartile of all samples for which the gene was not found on an amplicon (that is, the average expression of the unamplified gene). Weaken expressing genes (that is, with an average baseline FPKM upper quartile <5) were removed from the analysis. We then computed the fold change in expression of each gene on each amplicon as the FPKM upper quartile +1 of the amplified gene divided by the average FPKM upper quartile +1 of the unamplified samples, removing any fold changes that were 5 s.d. from the mean fold change and report the distribution of fold changes versus the CN. A Tukey's range test was used to test the significance between the slope of the FPKMs for circular and amplified noncircular. Transcript fusions for the TCGA samples were derived from the TumorFusions portal19-21. Fusion analysis was performed by taking the total number of fusions landing in an amplicon class divided by the total bp of all amplicons belonging to that amplicon class within the TCGA dataset to obtain an expected number of fusion events per bp for each amplicon class. To test for enrichment in circular amplicons compared to noncircular amplicons, a binomial test was performed by computing the probability of observing the total number of fusion events on circular amplicons, using the expected number of fusion events per bp of the noncircular amplicon class. ATAC-seq profiles were available for 24 samples22-24. The TCGA ATAC-seq data are provided as a count matrix, where each row is a peak (represented as hg38 coordinates) and each column is a TCGA sample. We remapped the peaks onto hg19 coordinates using LiftOver. We then intersected each ATAC-seq peak with the amplicons of the 36 samples. For each ATAC-seq peak that intersected with an amplicon, the CN normalized fold change in the ATAC-seq signal was computed as follows. For each sample, the normalized ATAC-seq signal was computed as the ATAC-seq signal of the sample for that peak divided by the estimated CN of that genomic region using the TCGA SNPs CN profile data. We then computed the CN normalized fold change divided by the normalized ATAC-seq signal of the sample with the intersecting amplicon divided by the mean normalized ATAC-seq signal of all samples without an amplicon intersecting with that peak. A Wilcoxon rank-sum test was used to test the significance between the two distributions.

Inferring breakpoint homologies. For each breakpoint, sequencing reads around ±1,000 bp of the breakpoint were locally reassembled with SvABA v.1.1.0 (ref. 41) to produce a contiguous consensus sequence of each breakpoint, precise breakpoint positions and the level of homology at breakpoints.

Chromothripsis analysis. Chromothripsis events were called with the ShatterSeek software v.0.4 (ref. 27) using somatic CN and structural variation calls as the input data. Structural variation clusters per patient were then defined as having or not having chromothripsis using the published set of statistical criteria, including correction for false discovery rate where applicable. We omitted the fragment joint test to relax test stringency and detect higher chromothripsis-like events to test positive association, if any, between chromothripsis and ecDNA regions. We defined patient having chromothripsis if ≥1 structural variation cluster had a chromothripsis event. A chi-squared test was used to evaluate positive enrichment of ecDNA and chromothripsis events at both locus- and patient-level. For locus-level enrichment, breakpoint regions for structural variation clusters and ecDNA region were overlapped using the bedtools intersect command v.2.29.0 (ref. 28).

Tandem duplicator phenotype (TDP). Tandem duplication calls from the TCGA and PCAWG were used to call TDP status using a method published previously29. The resulting sample-level TDP calls were then tallied with AmpliconArchitect-called ecDNA calls.

Statistical analysis. All data analyses were conducted in R v.3.3.2 and Python v.2.7.11 or v.3.5.4. Survival curves were estimated with the Kaplan–Meier method and comparison of survival curves between groups was performed with the log rank test in the R package ‘survival’ v.3.1.11. Hazard ratios were estimated with the Cox proportional-hazards regression model of the sample for the survival package. For further details, see the Nature Research Reporting Summary.

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

Data availability
Information on accessing the data from the ICGC, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. All open access TCGA data are publicly available through the National Cancer Institute Genomic Data Commons (https://gdac.cancer.gov/). The datasets marked 'Controlled' contain potentially identifiable information and require authorization from the ICGC and TCGA Data Access Committees. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access sequencing data, researchers need to apply to the TCGA Data Access Committee via the database of Genotypes and Phenotypes (https://dbGaP.ncbi.nlm.nih.gov/ca/wega.cgi?page=login) for access to the TCGA portion of the dataset and to the ICGC Data Access Compliance Office (http://igcc.dako) for the ICGC portion. All images analyzed are available from figshare at https://figshare.com/s/6c3e26edc1aba2999b20af20 and https://figshare.com/s/5ab21473a84aa3353391.

References
37. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell 33, 676–689.e3 (2018).
38. Hu, X. et al. TumorFusions: an integrative resource for cancer-associated transcript fusions. Nucleic Acids Res. 46, D1144–D1149 (2018).
39. Yoshihara, K. et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 34, 4845–4854 (2015).
40. Torres-Garcia, W. et al. PRADA: pipeline for RNA sequencing data analysis. Bioinformatics 30, 2224–2226 (2014).
41. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28, 581–591 (2018).
42. Quillinan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.1–11.12.14 (2014).

Acknowledgements
This work was supported by the Ludwig Institute for Cancer Research (P.S.M.), Defeat GBM Program of the National Brain Tumor Society (P.S.M.), NVIDIA Foundation, Compute for the Cure (P.S.M.), Ben and Catherine Ivy Foundation (P.S.M.), generous donations from the Ziering Family Foundation in memory of Sigi Ziering (P.S.M.) and Ruth L. Kirschstein National Research Service Award. This work was also supported by the following National Institutes of Health grants: NS73831 (to P.S.M.), GM114462 (to V.B.), R01 CA190121, R01 CA37208 and R21 NS114873. This work was supported by Cancer Center Support Grant P30 CA034196 (R.G.W.V.), grant nos. R53CA209919 (to H.Y.C.) and RM1-HG007735 (to H.Y.C.), R35GM136300 (to C.R.B.), National Science Foundation grant nos. NSF-IIS-1318836 and NSF-D15-1458557 (to V.B.) and grants from the Musella Foundation, B*CURED Foundation, Brain Tumor Charity and Department of Defense grant nos W81XWH1910246 (to R.G.W.V). H.Y.C. is an Investigator of the Howard Hughes Medical Institute. The results published in this paper are in whole or part based on data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and the International Cancer Genome Consortium (https://icgc.org/). Analysis of the TCGA and International Cancer Genome Consortium datasets was made possible through the Cancer Genomics Cloud of the Institute for Systems Biology (ISB-ICGC) and the Amazon Web Services Cloud, respectively.

Author contributions
H.K., N.P.N., P.S.M., V.B. and R.G.W.V. conceived the study and designed the experiments. Data analysis was led by H.K. and N.P.N. in collaboration with S.W.J., L. Ruebeck, V.D., S.N., E.M., U.R., H.Y.C., E.Y. and C.R.B. Cloud data access was performed by H.K. and S.N. The FISH experiments were performed by K.T., S.W., E.Y. and A.D.G. EcSeg was performed by U.R. and J. Liu. The CIRCLE-seq data were provided by I.H.S. and A.G.H. K., N.P.N., P.S.M., V.B. and R.G.W.V. wrote the manuscript. E.Y. reviewed the manuscript. All coauthors discussed the results and commented on the manuscript and the supplementary information.

Competing interests
H.Y.C., P.S.M., V.B. and R.G.W.V. are scientific cofounders of Boundless Bio and serve as consultants. V.B. is a cofounder and has equity interest in Digital Proteomics, and receives income from Digital Proteomics. 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. N.P.N. and K.T. are employees of Boundless Bio.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0678-2. Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0678-2. Correspondence and requests for materials should be addressed to P.S.M., V.B. or R.G.W.V.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Amplicon classification. a. Validation on cell line data. Validation of the classification scheme on cell line data with FISH experiments for detecting ecDNA from the Turner et al. and deCarvalho et al. studies, in addition to newly generated data. FISH probes were designed for selected oncogenes and DAPI staining was performed to determine whether the FISH probe landed on chromosomal DNA or ecDNA. For each cell (represented as an image of the cell in metaphase), the number of positive ecDNA probes were counted, and for each cell line, the average positive ecDNA per cell was reported. For each probe, we report whether it landed in an amplicon (inferred from AmpliconArchitect), and if so, what was the amplicon’s classification. The distribution for the average ecDNA per cell between the Circular and non-circular classes was statistically significantly different (p-value < 1e-9; Wilcoxon rank sum test). b–d. Whole-genome sequencing derived based Circular amplicon regions (blue) were validated with Circle-seq (red) for three neuroblastoma samples (CB2001, CB2022, and CB2050, respectively) used in the Koche et al. study.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Circular vs amplified non-circular amplification comparisons. a. 24 recurrently amplified oncogenes significantly overlap circular regions (z-score 37.8), especially compared to amplified non-circular regions (z-scores of 30.4, 29.5, 28.0 for Linear, Heavily-rearranged, and BFB). b. For all oncogenes on amplicons with copy number \( \geq 4 \) and present in at least 5 samples across the cohort, we show the class distribution of that oncogene. The oncogenes are ordered by proportion on circular amplification. c. For the 24 recurrent oncogenes known to be activated via amplification (Zack et al. Nat Gen. 2013), we report the average copy number for the oncogenes for circular amplification versus amplified-noncircular amplification. d. Breakpoint location across all samples for each recurrently amplified oncogene. We identified all breakpoints from each sample containing the recurrent oncogene on ecDNA and report the total number of breakpoints across this region in 1kb binned windows. e. Distribution of breakpoint locations across all circular samples for each recurrently amplified oncogene. We identified all breakpoints from each sample containing the recurrent oncogene on ecDNA. Shown is the distribution of the number of breakpoints in each bin, which closely follows a Poisson distribution, suggesting that the breakpoints are mostly randomly distributed across the region.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Genome instability vs amplicon classes. a. Chromosome arm aneuploidy scores showing no or marginal difference in chromosomal arm level events between circular and non-circular amplification classes. b. Genome doubling events by amplification class. c. Distribution for total DNA loss segments by amplification class. WGS-inferred CNV data was used to count the total number of DNA losses within a sample. A DNA loss was defined as a segment with CN < 2. d. Distribution for total DNA gain segments by amplification class. WGS-inferred CNV data was used to count the total number of DNA gains within a sample. A DNA gain was defined as a segment with CN > 2. Circular samples contain statistically significantly more DNA gains than BFB, Heavily-rearranged, Linear, and No-fSCNA (p-value < 0.03, <0.03, <1e-20, and <1e-111, respectively; Wilcox Rank Sum Test). e. Breakpoint homology by amplification class. f. Comparison of amplicon versus locus-level chromothripsis (Pearson’s Chi-squared test data: X-squared = 4674.7, df = 3, p-value < 2.2e-16). g. Comparison of sample category versus sample-level chromothripsis (Pearson’s Chi-squared test data: X-squared = 21.58, df = 3, p-value 8e-05 (excludes ‘No fSCNA detected’ category)). h. Comparison of sample category versus sample-level tandem duplication (Pearson’s Chi-squared test data: X-squared = 7.39, df = 3, p-value 0.06 (excludes ‘No fSCNA detected’ category)).
Extended Data Fig. 4 | Gene expression of amplicon classes. Copy number of the oncogene versus its fold-change in FPKM for all oncogenes with a copy count greater than 4, for each oncogene on each amplicon. The fold-change in FPKM is computed as the oncogene’s (FPKM-UQ+1) divided by the average of (FPKM-UQ+1) for the same oncogene in all other tumor samples from the same cohort for which the oncogene is not on any amplicon (that is, not amplified). Linear regression lines, using fold change = m*CNV+b where m and b are selected to minimize error of the fit, are shown for each class. Tukey’s range test shows oncogenes on circular structures are significantly different to oncogenes on non-circular structures (p-value < 1e-7).
Extended Data Fig. 5 | Lymph node stage vs amplicon classes. Lymph node stage for primary tumors showing samples with amplification are more likely to have spread to the lymph node at time of diagnosis (Chi-square test; df=4; p-value<1e−05).
Extended Data Fig. 6 | Cell cycle and immune infiltrate gene expression signatures vs amplicon classes. a. Cell Cycle gene expression signature single sample GSEA (ssGSEA) scores by amplification category. b. Immune infiltrate gene expression signature single sample GSEA (ssGSEA) scores by amplification category.
# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](https://www.nature.com/nature/research/policies/editorial-policies) and the [Editorial Policy Checklist](https://www.nature.com/nature/research/policies/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.

| Item                                                                 | 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](https://www.nature.com/nature/research/articles/using-statistics-in-biological-research) contains articles on many of the points above.*

## Software and code

Policy information about [availability of computer code](https://www.nature.com/nature/research/policies/software)

**Data collection**

Every sliced TCGA WGS BAM was prepared by using the following command: `java -Xmx12g -jar /gatk/gatk-package-4.0.12.0-local.jar PrintReads -I INPUT.bam -O output.sliced.bam -R Homo_sapiens_assembly19.fasta -L target.gatk.interval_list`

**Data analysis**

AmpliconArchitect (no version information) is available at [https://github.com/virajbdeshpande/AmpliconArchitect](https://github.com/virajbdeshpande/AmpliconArchitect)

ecSeg (no version information) is available at [https://github.com/UCRajkumar/ecSeg](https://github.com/UCRajkumar/ecSeg)

GATK version: version 4.0.12.0

SvABA: version 134

R: version 3.3.2

Python: version 2.7.11 and 3.5.4

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 Research [guidelines for submitting code & software](https://www.nature.com/nature/research/policies/software) for further information.

## Data

Policy information about [availability of data](https://www.nature.com/nature/research/policies/software)

All manuscripts must include a [data availability statement](https://www.nature.com/nature/research/policies/software). This statement should provide the following information, where applicable:

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

All data analyzed in the manuscript were obtained via The Cancer Genome Atlas and International Cancer Genome Consortium both of which are publicly available.
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
  - No statistical methods were used to predetermine sample size.

- Data exclusions
  - All publicly available whole genome sequencing data from The Cancer Genome Atlas and International Cancer Genome Consortium were analyzed. No samples were excluded a priori.

- Replication
  - Replication was not performed.

- Randomization
  - No randomization was done to allocate samples into experimental groups.

- Blinding
  - The samples we analyzed were deidentified by The Cancer Genome Atlas and International Cancer Genome Consortium.

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
  - Human research participants
  - Clinical data
  - Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics

The samples consist of 3212 tumors and 1810 matched normals from 3212 unique patients. They include 83 patients from B-cell lymphoma, 11 patients from Biliary tract, 112 patients from Bladder, 159 patients from Breast, 66 patients from Cervical, 74 patients from Colorectal, 112 patients from Esophageal, 37 patients from Ewing Sarcoma, 145 patients from Gastric, 47 patients from Glioblastoma, 153 patients from Head and Neck, 252 patients from Liver, 85 patients from Lower Grade Glioma, 143 patients from Lung Adeno, 50 patients from Lung Squamous cell, 95 patients from Lymphoid leukemia, 30 patients from Myeloid Disorders, 48 patients from Myeloid leukemia, 11 patients from Oral, 70 patients from Ovarian, 213 patients from Pancreatic, 188 patients from Pediatric Brain, 301 patients from Prostate, 198 patients from Renal, 36 patients from Sarcoma, 164 patients from Skin, 136 patients from Thyroid papillary, 143 patients from Uterine Corpus Endometrial, and 50 patients from Uveal melanoma.

Recruitment

All recruitment was done by The Cancer Genome Atlas and International Cancer Genome Consortium.

Ethics oversight

NIH, The Cancer Genome Atlas, International Cancer Genome Consortium

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