LETTER

doi:10.1038/nature21356

Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity

Kristen M. Turner[1]*, Viraj Deshpande[2]*, Doruk Beyter[2]*, Tomoyuki Koga[1], Jessica Rusert[3], Catherine Lee[3], Bin Li[1], Karen Arden[1], Bing Ren[1], David A. Nathanson[4], Harley I. Kornblum[4,5], Michael D. Taylor[6], Sharmeela Kaushal[7], Webster K. Cavenee[1], Robert Wechsler-Reya[3], Frank B. Furnari[1], Scott R. Vandenberg[8], P. Nagesh Rao[9], Geoffrey M. Wahl[10]§, Vineet Bafna[2]§ & Paul S. Mischel[1,7,11]§

Human cells have twenty-three pairs of chromosomes. In cancer, however, genes can be amplified in chromosomes or in circular extrachromosomal DNA (ecDNA), although the frequency and functional importance of ecDNA are not understood[1–4]. We performed whole-genome sequencing, structural modelling and cytogenetic analyses of 17 different cancer types, including analysis of the structure and function of chromosomes during metaphase of 2,572 dividing cells, and developed a software package called ECdetect to conduct unbiased, integrated ecDNA detection and analysis. Here we show that ecDNA was found in nearly half of human cancers; its frequency varied by tumour type, but it was almost never found in normal cells. Driver oncogenes were amplified most commonly in ecDNA, thereby increasing transcript level. Mathematical modelling predicted that ecDNA amplification would increase oncogene copy number and intratumoural heterogeneity more effectively than chromosomal amplification. We validated these predictions by quantitative analyses of cancer samples. The results presented here suggest that ecDNA contributes to accelerated evolution in cancer.

Cancers evolve in rapidly changing environments from single cells into genetically heterogeneous masses. Darwinian evolution selects for survival of the fittest cells, that is, those that are best suited to their environment. Heterogeneity provides a pool of mutations upon which selection can act[1,5–9]. Cells that acquire fitness-enhancing mutations are more likely to pass these mutations on to daughter cells, driving neoplastic progression and therapeutic resistance[10,11]. One common type of cancer mutation, oncogene amplification, can be found either in chromosomes or in nuclear ecDNA elements, including double minutes[2–4,12–14]. Relative to chromosomal amplicons, ecDNA is less stable, segregating unequally to daughter cells[15,16]. Double minutes are reported to occur in 1.4% of cancers with a maximum of 31.7% in neuroblastoma, based on the Mitelman database[4,17]. However, the scope of ecDNA in cancer has not been accurately quantified, the oncogenes contained therein have not been systematically examined and the impact of ecDNA on tumour evolution has yet to be determined.

DNA sequencing permits unbiased analysis of cancer genomes, but it cannot spatially resolve amplicons to specific chromosomal or extrachromosomal regions. Bioinformatic analyses can potentially infer DNA circularity[18], but the number of extrachromosomal amplicons may vary from cell to cell. Consequently, copies of oncogenes amplified on ecDNA may be greatly underestimated. Cytogenetic analysis of tumour cells during metaphase can localize amplicons, but this technique does not permit unbiased analysis. To quantify the spectrum of ecDNA in human cancer cells and systematically analyse the contents of the ecDNA, we integrated whole-genome sequencing of 117 cancer cell lines, patient-derived tumour cell cultures and tumour tissues from a range of cancer types (Fig. 1a) with bioinformatic and cytogenetic analysis of 2,049 cells in metaphase from 72 cancer cell samples for which cells during metaphase could be obtained. Additionally, 290 cells in metaphase from 10 immortalized cell cultures, and 233 cells in metaphase from 8 normal tissue cultures were analysed, with a total of 2,572 cells in metaphase analysed (Source Data of Fig. 1 and Methods).

The fluorescent dye DAPI (4′,6-diamidino-2-phenylindole) allows ecDNA detection (Fig. 1b), which was confirmed using genomic DNA and centromeric FISH (fluorescence *in situ* hybridization) probes (Fig. 1b–d and Extended Data Fig. 1). We developed an image analysis software package called ECdetect (Fig. 1e and Methods), providing a robust, reproducible and highly accurate method for quantifying ecDNA from DAPI-stained metaphases in an unbiased, semi-automated fashion. ECdetect accurately detected ecDNA and this detection rate was highly correlated with visual detection ($r = 0.98$, $P < 2.2 \times 10^{-16}$; Fig. 1f), allowing the quantification of 2,572 cells in metaphase, including at least 20 cells in metaphase from each sample.

ecDNA was abundant in the cancer samples (Fig. 2a), but was rarely found in normal cells. Approximately 30% of the ecDNA were paired double minutes (Source Data of Fig. 2). ecDNA levels varied among tumour types, with substantially higher levels in patient-derived cultures (Fig. 2b). Using the conservative metric of at least two ecDNA copies in ≥10% (2 out of 20) cells in metaphase, ecDNA was detected in nearly 40% of tumour cell lines and nearly 90% of patient-derived brain tumour models (Fig. 2c, d, Extended Data Fig. 2 and Methods). No significant associations between ecDNA level and primary tumour or metastatic status; untreated or treated samples; or un-irradiated or post-irradiated tumours were detected (Source Data of Fig. 2). The diverse array of treatments relative to the sample size limited our ability to conclusively determine the effect of specific therapies on ecDNA levels. ecDNA number varied greatly from cell to cell within a tumour culture (Fig. 2e–g, Extended Data Fig. 3 and Supplementary Information 2.3), as quantified by the Shannon diversity index[19]. These data demonstrate that ecDNA is common in cancer cells, varies greatly from cell to cell and is very rare in cells derived from normal tissue.

[1]Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, California 92093, USA. [2]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, California 92093, USA. [3]Tumor Initiation and Maintenance Program, NCI-Designated Cancer Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California 92037, USA. [4]Department of Molecular and Medical Pharmacology, David Geffen UCLA School of Medicine, Los Angeles, California 90095, USA. [5]Neuropsychiatric Institute–Semel Institute for Neuroscience and Human Behavior and Department of Psychiatry and Biobehavioral Sciences, David Geffen UCLA School of Medicine, Los Angeles, California 90095, USA. [6]The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. [7]Moores Cancer Center, University of California at San Diego, La Jolla, California 92093, USA. [8]Department of Pathology, University of California San Francisco, San Francisco, California 94143, USA. [9]Department of Pathology and Laboratory Medicine, David Geffen UCLA School of Medicine, Los Angeles, California 90095, USA. [10]Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA. [11]Department of Pathology, University of California at San Diego, La Jolla, California 92093, USA.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

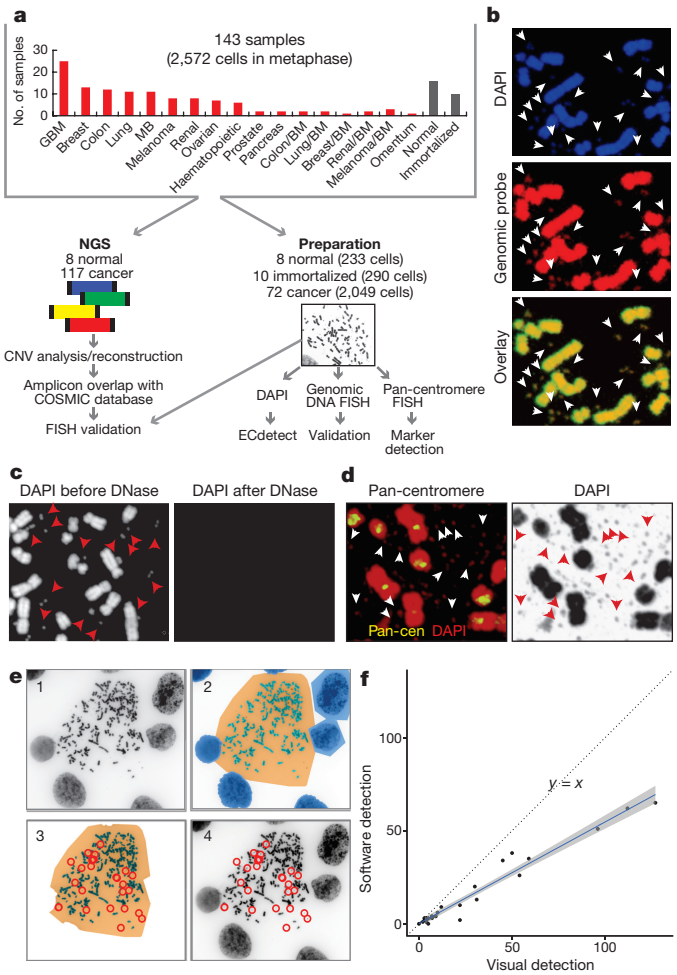


Figure 1 | Integrated next-generation DNA sequencing and cytogenetic analysis of ecDNA. a, Schematic diagram of experimental flow. BM, brain metastasis; GBM, glioblastoma; MB, medulloblastoma. **b**, Representative cells during metaphase stained with DAPI and a genomic DNA FISH probe (ecDNA, arrows). **c**, DNase treatment abolishes DAPI staining of chromosomal and ecDNA (arrows). **d**, Pan-centromeric FISH shows that a centromere in the ecDNA is absent (arrows). **e**, Schematic illustration of ECdetect. (1) DAPI-stained metaphase as input, (2) semi-automated identification of ecDNA search region through segmentation, (3) conservative filtering, removing non-ecDNA components and (4) ecDNA detection and visualization. **f**, Pearson correlation between software-detected and manual calls of ecDNA ($r = 0.98$, $P < 2.2 \times 10^{-16}$).

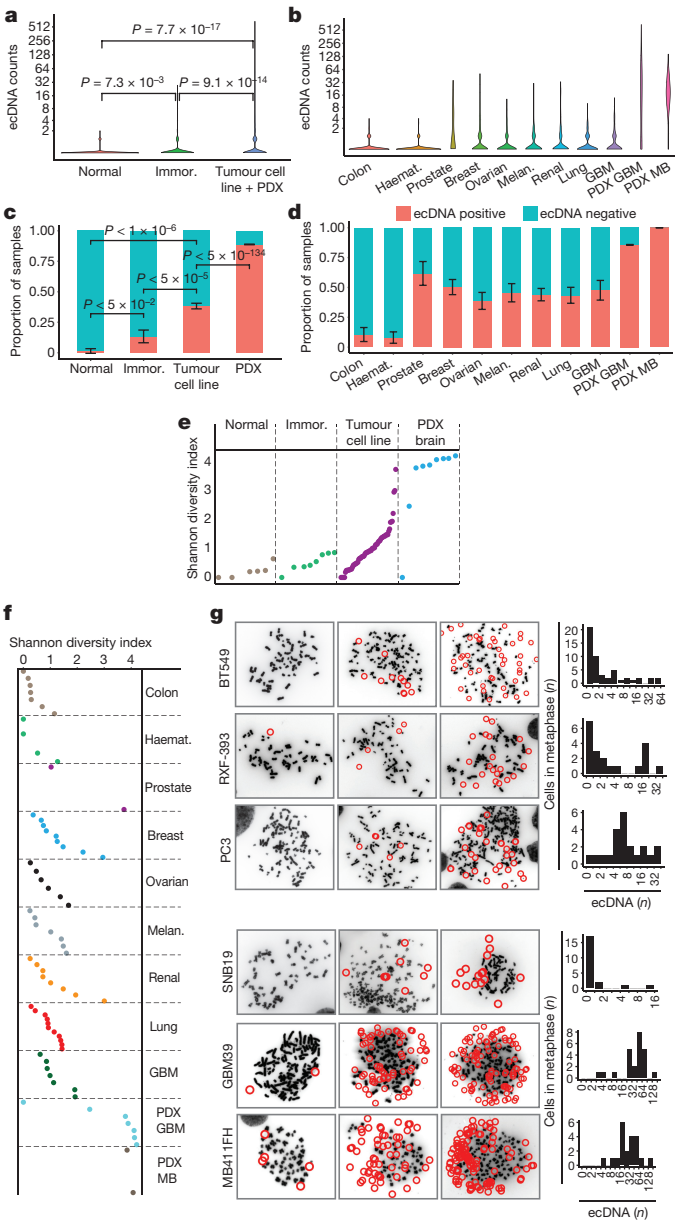


Figure 2 | ecDNA is found in nearly half of cancers and contributes to intratumoural heterogeneity. a, Distribution of ecDNA elements per cell in metaphase from 72 cancer, 10 immortalized and 8 normal cell cultures, Wilcoxon rank-sum test. PDX, patient-derived xenograft. **b**, ecDNA distribution per cell in metaphase stratified by tumour type. **c**, Proportion of samples with two or more ecDNA elements in at least two out of 20 cells (positive for ecDNA) in metaphase. Data shown as mean ± s.e.m. (Methods). **d**, Proportion of tumour cultures positive for ecDNA by tumour type. **e**, Shannon diversity index. Each dot represents an individual cell line sampled with ≥ 20 cells in metaphase. **f**, Shannon diversity index by tumour type. **g**, DAPI-stained cells in metaphase of cell lines with histograms.

Whole-genome sequencing with a median coverage of 1.19× (Extended Data Fig. 4) showed focal amplifications that were nearly identical to the amplifications found in The Cancer Genome Atlas (TCGA) analyses of the same cancer types (Fig. 3a and Source Data of Fig. 3), including amplified oncogenes found in a pan-cancer analysis of 13 different cancer types[20]. All of the amplified oncogenes tested were found solely in the ecDNA, or concurrently in ecDNA and chromosomal homogenous staining regions (HSRs) (Fig. 3b, c and Extended Data Figs 5, 6). Oncogenes amplified in ecDNA showed high expression levels of mRNA transcripts (Fig. 3d) and the copy-number diversity of commonly amplified oncogenes in ecDNA far exceeded oncogene copy-number diversity if the oncogenes were located on other chromosomal loci (Extended Data Fig. 7).

To determine whether extra- and intrachromosomal structures had a common origin, we developed 'AmpliconArchitect' to elucidate the finer genomic structure using sequencing data (Methods). To better understand the relationship between subnuclear location and amplicon structure, we took advantage of a spontaneously occurring subclone of GBM39 cells in which a high copy *EGFR* mutant, EGFRvIII (an *EGFR* mutant with exons 2–7 deleted), shifted from the ecDNA exclusively to HSRs. Independent replicates of GBM39 containing an ecDNA amplicon, showed a consistent circular structure of 1.29 Mb containing one copy of EGFRvIII (Extended Data Fig. 8). Notably, the GBM39 subclone containing EGFRvIII exclusively on HSRs had an identical structure with tandem duplications containing multiple copies of EGFRvIII, indicating that the HSRs arose from reintegration of the EGFRvIII-containing ecDNA elements[14] (Extended Data Fig. 8). In GBM39 cells, resistance to EGFR tyrosine kinase inhibitors is caused by reversible loss of EGFRvIII from ecDNA[21]. Structural analysis

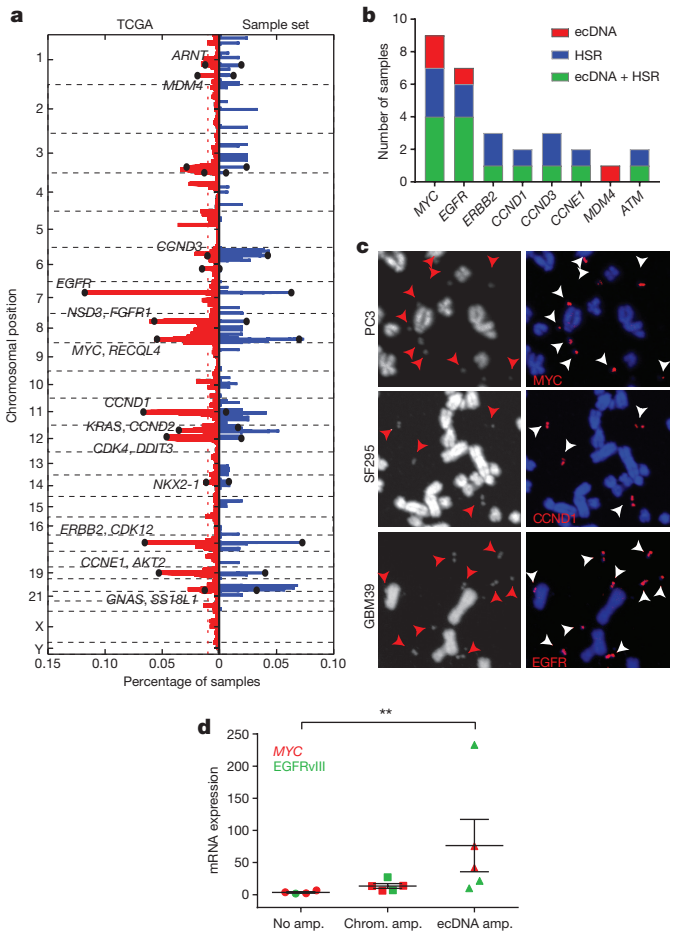


Figure 3 | The most common focal amplifications in cancer are contained on ecDNA. a, Comparison of the frequency of focal amplifications detected by next generation sequencing of the 117 cancer samples studied here (blue) with those of matched tumour types in the TCGA (red) demonstrates significant overlap and representative sampling (P value 10^{-6} based upon random permutations of TCGA amplicons; Methods). **b**, Localization of oncogenes by FISH. **c**, Representative FISH images of focal amplifications on ecDNA (arrows). **d**, EGFRvIII and *MYC* mRNA level, measured by qPCR ($P < 0.001$, Mann–Whitney U-test). Data are mean ± s.e.m.; $n = 17$; each data point represents an average qPCR value of three technical replicates.

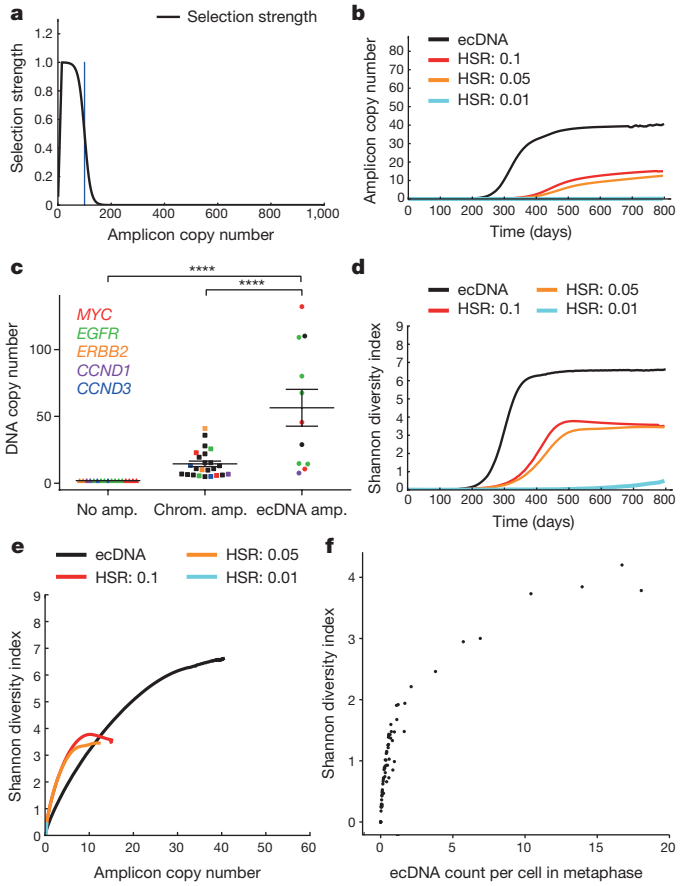


Figure 4 | Theoretical model for focal amplification via extrachromosomal and intrachromosomal mechanisms. Simulated change in copy number via random segregation (ecDNA) or mitotic recombination (HSR), starting with 10^5 cells, 100 of which carry amplifications. **a**, The selection function $f_{100}(k)$ reaches a maximum for $k = 15$, then decays logistically. **b**, Growth in amplicon copy number over time. **c**, DNA copy number stratified by oncogene location. ($P < 0.001$, ANOVA/Tukey's multiple comparison). $n = 52$; data points include top five amplified oncogenes, mean ± s.e.m. **d**, Change in heterogeneity (Shannon diversity index) over time. **e**, Correlation between copy number and heterogeneity. **f**, Experimental data showing correlation between ecDNA counts and heterogeneity matches the simulation in **e**.

revealed a conservation of the fine structure of the EGFRvIII amplicon containing ecDNA in naive cells, during treatment and upon regrowth after discontinuation of therapy (Extended Data Fig. 9), indicating that ecDNA can dynamically relocate to chromosomal HSRs while maintaining key structural features[14,22].

We next investigated whether ecDNA localization conferred a particular benefit, relative to chromosomal amplification. We hypothesized that ecDNA amplification may enable an oncogene to rapidly reach higher copy number because of the unequal segregation to daughter cells[15] than would be possible by intrachromosomal amplification. We used a simplified Galton–Watson branching process to model the evolution of a tumour[23], where each cell in the current generation either replicates or dies to create the next generation. A cell with k copies of the amplicon is selected for replication with probability b_k, as defined by $\frac{b_k}{(1-b_k)} = 1 + sf_m(k)$. We provided a positive selection bias towards cells with higher ecDNA counts by choosing s in the range of 0.5 to 1, and different selection regimes for f. Specifically, $f_m(k)$ increases to a maximum value $f_m(15) = 1$, then declines in a logistic manner with $f_m(m) = 0.5$ to reflect metabolic constraints (Methods). We allowed the amplicon copy number to grow to 1,000 copies (Extended Data Fig. 10), but set $b_k = 0$ for $k \geq 10^3$. During cell division, the $2k$ copies resulting from the replication of each of the k ecDNA copies segregate independently into the two daughter cells. We contrasted this with an intrachromosomal model of duplication with identical selection constraints, but with the change in copy number affected by mitotic recombination, and achieved by increasing or decreasing k by 1, with duplication probability P_d. A range of values for P_d, ($0.01 \leq P_d \leq 0.1$) was used, where the upper boundary reflects a change in copy number once every five divisions. The full assumptions of the model are explained in detail in Supplementary Information 4. Starting with an initial population of 10^5 cells, with $s = 0.5$, $m = 100$ and a selection function $f_{100}(k)$ (Fig. 4a), we find that an oncogene can reach a much higher copy number in a tumour if it is amplified on ecDNA, rather than on a chromosome (Fig. 4b). As predicted by the model, we detected a significantly higher copy number of the most frequently amplified oncogenes *EGFR* (including EGFRvIII) and *MYC*, when they were contained within ecDNA instead of within chromosomes (Fig. 4c). We also reasoned that if an oncogene is amplified intrachromosomally, the heterogeneity of the tumour (in terms of the distribution of copies of the oncogene) would stabilize at a much lower level. By contrast, unequal segregation of ecDNA would probably rapidly enhance heterogeneity and maintain it. Our model consistently confirmed this prediction (Fig. 4d) for a wide range of simulation parameters (Supplementary Information 4.3). The heterogeneity of

copy-number change stabilizes and even decreases over time[10,24], much as predicted in Fig. 4d. We also tested the validity of the model by comparing the Shannon diversity index against the average number of amplicons per cell in our tumour samples. Heterogeneity of a tumour with respect to oncogene copy number would be more likely to rise relatively slowly if it is present on a chromosome, but would rise more rapidly and be maintained much longer, if that oncogene is present on ecDNA, as confirmed by a plot of Shannon diversity index versus copy number (Fig. 4e). Moreover, the predicted correlation in Fig. 4e is completely recapitulated by the experimental data (Fig. 4f), thereby validating the central tenets of the model.

There is growing evidence that genetically heterogeneous tumours are remarkably difficult to treat[10]. The data presented here identifies a mechanism by which tumours maintain cell-to-cell variability in the copy number and transcriptional level of oncogenes that drive tumour progression and drug resistance. We suggest that extrachromosomal oncogene amplification may enable tumours to adapt more effectively to variable environmental conditions by increasing the likelihood that a subpopulation of cells will express that oncogene at a level that maximizes tumour proliferation and survival[12,21,25–28], rendering tumours progressively more aggressive and difficult to treat over time. Even when using a selection function that only mildly depends on copy number, we detected a very large difference between intra- and extrachromosomal amplification mechanisms leading to a higher copy number of amplicons and greater heterogeneity in copy number. Thus, even small increases in selection advantage conferred by oncogenes amplified on ecDNA would be expected to yield a very high fitness advantage (Supplementary Information 4.3). The notably high frequency of ecDNA in cancer, as shown here, coupled to the benefits to tumours of extrachromosomal gene amplification relative to chromosomal inheritance, suggest that oncogene amplification on ecDNA may be a driving force in tumour evolution and the development of genetic heterogeneity in human cancer. Understanding the underlying molecular mechanisms of tumour evolution, including oncogene amplification in ecDNA, may help to identify more effective treatments that either prevent cancer progression or more effectively eradicate tumours.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 August; accepted 23 December 2016.
Published online 8 February 2017.

1. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
2. Stark, G. R., Debatisse, M., Giulotto, E. & Wahl, G. M. Recent progress in understanding mechanisms of mammalian DNA amplification. *Cell* **57,** 901–908 (1989).
3. Schimke, R. T. Gene amplification in cultured animal cells. *Cell* **37,** 705–713 (1984).
4. Fan, Y. *et al.* Frequency of double minute chromosomes and combined cytogenetic abnormalities and their characteristics. *J. Appl. Genet.* **52,** 53–59 (2011).
5. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
6. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27,** 15–26 (2015).
7. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12,** 323–334 (2012).
8. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13,** 795–806 (2012).
9. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
10. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2016).
11. Gillies, R. J., Verduzco, D. & Gatenby, R. A. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer* **12,** 487–493 (2012).
12. Von Hoff, D. D., Needham-VanDevanter, D. R., Yucel, J., Windle, B. E. & Wahl, G. M. Amplified human MYC oncogenes localized to replicating submicroscopic circular DNA molecules. *Proc. Natl Acad. Sci. USA* **85,** 4804–4808 (1988).
13. Garsed, D. W. *et al.* The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26,** 653–667 (2014).
14. Carroll, S. M. *et al.* Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Mol. Cell. Biol.* **8,** 1525–1533 (1988).
15. Windle, B., Draper, B. W., Yin, Y. X., O'Gorman, S. & Wahl, G. M. A central role for chromosome breakage in gene amplification, deletion formation, and amplicon integration. *Genes Dev.* **5,** 160–174 (1991).
16. Kanda, T., Otter, M. & Wahl, G. M. Mitotic segregation of viral and cellular acentric extrachromosomal molecules by chromosome tethering. *J. Cell Sci.* **114,** 49–58 (2001).
17. Mitelman, F., Johansson, B. & Mertens, F. *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.* http://cgap.nci.nih.gov/Chromosomes/Mitelman (2016).
18. Sanborn, J. Z. *et al.* Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. *Cancer Res.* **73,** 6036–6045 (2013).
19. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Reports* **6,** 514–527 (2014).
20. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
21. Nathanson, D. A. *et al.* Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* **343,** 72–76 (2014).
22. Storlazzi, C. T. *et al.* Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res.* **20,** 1198–1206 (2010).
23. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107,** 18545–18550 (2010).
24. Li, X. *et al.* Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett's esophagus. *Cancer Prev. Res.* **7,** 114–127 (2014).
25. Mishra, S. & Whetstine, J. R. Different facets of copy number changes: permanent, transient, and adaptive. *Mol. Cell. Biol.* **36,** 1050–1063 (2016).
26. Schimke, R. T., Kaufman, R. J., Alt, F. W. & Kellems, R. F. Gene amplification and drug resistance in cultured murine cells. *Science* **202,** 1051–1055 (1978).
27. Nikolaev, S. *et al.* Extrachromosomal driver mutations in glioblastoma and low-grade glioma. *Nat. Commun.* **5,** 5690 (2014).
28. Biedler, J. L., Schrecker, A. W. & Hutchison, D. J. Selection of chromosomal variant in amethopterin-resistant sublines of leukemia L1210 with increased levels of dihydrofolate reductase. *J. Natl. Cancer Inst.* **31,** 575–601 (1963).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Kolodner, W. Mischel, D. Geschwind, members of the Mischel laboratory, A. Akbari, A. Iranmehr and A. Patel for helpful comments. This work was supported by the Ludwig Institute for Cancer Research (P.S.M., B.R., K.A., W.K.C., F.B.F.), Defeat GBM Program of the National Brain Tumor Society (P.S.M., F.B.F.), The Ben and Catherine Ivy Foundation (P.S.M.), generous donations from the Ziering Family Foundation in memory of Sigi Ziering (P.S.M.); The Susan G. Komen Foundation (SAC110036), The Leona M. and Harry B. Helmsley Charitable Trust (2012-PG-MED002) and The Breast Cancer Research Foundation (BCRF) to G.M.W.; CureSearch for Children's Cancer and a Leadership Award from the California Institute for Regenerative Medicine to R.W.R. This work was also supported by the following NIH grants: NS73831 (P.S.M.), GM114362 (V.B., V.D., D.B.), NS80939 (F.B.F.), CA014195 and CA159859 (G.M.W.) and CA151819 (D.A.N.) and T32CA121938 (K.M.T.) and NSF grants: NSF-IIS-1318386 and NSF-DBI-1458557 (V.B., V.D., D.B.).

Author Contributions K.M.T., V.D., D.B., V.B. and P.S.M. conceived and designed the study. K.M.T., V.D., D.B., T.K., J.R., C.L. and B.L. performed experiments. D.B., V.D. and V.B. developed the ECdetect software and performed mathematical modelling and simulations. V.D. and V.B. developed the AmpliconArchitect software. K.A., B.R., D.A.N., F.B.F., W.K.C., P.N.R. and G.M.W. provided analytic support. H.I.K., M.D.T., S.K, R.W.-R. and S.R.V. provided additional clinical samples and analytic support. K.M.T., V.D., D.B., V.B. and P.S.M. wrote the manuscript with feedback from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.S.M (pmischel@ucsd.edu) and for computational methods and tools to V.B. (vbafna@cs.ucsd.edu).

Reviewer Information *Nature* thanks C. Maley, A. Papenfuss and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cytogenetics. Metaphase cells were obtained by treating cells with Karyomax (Gibco) at a final concentration of 0.01 μg ml^{-1} for 1–3 h. Cells were collected, washed in PBS, and resuspended in 0.075 M KCl for 15–30 min. Carnoy's fixative (3:1 methanol:glacial acetic acid) was added dropwise to stop the reaction. Cells were washed an additional three times with Carnoy's fixative, before being dropped onto humidified glass sides for metaphase cell preparations. For ECdetect analyses, DAPI was added to the slides. Images in the main figures were captured with an Olympus FV1000 confocal microscope. All other images were captured at a magnification of 1,000× with an Olympus BX43 microscope equipped with a QiClick cooled camera. FISH was performed by adding the appropriate DNA FISH probe onto the fixed metaphase spreads. A coverslip was added and sealed with rubber cement. DNA denaturation was carried out at 75 °C for 3–5 min and the slides were allowed to hybridize overnight at 37 °C in a humidified chamber. Slides were subsequently washed in 0.4× SSC at 50 °C for 2 min, followed by a final wash in 2× SSC containing 0.05% Tween-20. Metaphase cells and interphase nuclei were counterstained with DAPI, a coverslip was applied and images were captured.

Cell culture. The NCI-60 cell line panel (gift from A. Shiau, obtained from NCI) was grown in RPMI-1640 with 10% FBS under standard culture conditions. Cell lines were not authenticated, as they were obtained from the NCI. The PDX cell lines were cultured in DMEM/F-12 medium supplemented with glutamax, B27, EGF, FGF and heparin. Lymphoblastoid cells (gift from B. Ren) were grown in RPMI-1640, supplemented with 2 mM glutamine and 15% FBS. IMR90 and ALS6-Kin4 (gift from J. Ravits and D. Cleveland) cells were grown in DMEM/F-12 supplemented with 20% FBS. Normal human astrocytes (NHA) and normal human dermal fibroblasts (NHDF) were obtained from Lonza and cultured according to Lonza-specific recommendations. Cell lines were not tested for mycoplasma contamination.

Tissue samples. Tissues were obtained from the Moores Cancer Center Biorepository Tissue Shared Resource with IRB approval (#090401). All samples were de-identified and patient consent was obtained. Additional tissue samples that were obtained were approved by the UCSD IRB (#120920).

DNA library preparation. DNA was sonicated to produce 300–500 bp fragments. DNA end repair was performed using End-it (Epicentre), DNA library adapters (Illumina) were ligated and the DNA libraries were amplified. Paired-end next-generation sequencing was performed and samples were run on the Illumina Hi-Seq using 100 cycles.

DNA extraction. Cells were collected and washed with 1 × cold PBS. Cell pellets were resuspended in buffer 1 (50 mM Tris pH 7.5, 10 mM EDTA, 50 μg ml^{-1} RNase A), and incubated in buffer 2 (1.2% SDS) for 5 min on ice. DNA was acidified by the addition of buffer 3 (3 M CsCl, 1 M potassium acetate, 0.67 M acetic acid) and incubated for 15 min on ice. Samples were centrifuged at 14,000*g* for 15 min at 4 °C. The supernatant was added to a Qiagen column and briefly centrifuged. The column was washed (60% ethanol, 10 mM Tris pH 7.5, 50 μM EDTA, 80 mM potassium acetate) and eluted in water.

DNase treatment. Metaphase cells were dropped onto slides and visualized with DAPI. Coverslips were removed and slides washed in 2 × SSC, and subsequently treated with 2.5% trypsin, and incubated at 25 °C for 3 min. Slides were then washed in 2 × SSC, DNase solution (1 mg ml^{-1}) was applied to the slide and cells were incubated at 37 °C for 3 h. Slides were washed in 2 × SSC and DAPI was again applied to the slide to visualize DNA.

ecDNA count statistics. In Fig. 2a, b the violin plots represent the distribution of ecDNA counts in different sample types. In order to compare the ecDNA counts between the different samples, we use a one-sided Wilcoxon rank-sum test, where the null hypothesis assumes that the mean ecDNA-count ranks of the compared sample types are equal.

Estimation of frequency of samples containing ecDNA. There is a wide variation in the number of ecDNA across different samples and within metaphases of the same sample. We want to estimate and compare the frequency of samples containing ecDNA for each sample type. We label a sample as being ecDNA positive by using the pathology standard: a sample is deemed to be ecDNA positive if we observe ≥ 2 ecDNA in ≥ 2 out of 20 metaphase images. Therefore, we ensure that every sample contains at least 20 metaphases.

We define indicator variable $X_{i,j} = 1$ if metaphase image j in sample i has ≥2 ecDNA elements, $X_{i,j} = 0$ otherwise. Let n_i be the number of metaphase images acquired for sample i. We assume that $X_{i,j}$ is the outcome of the jth Bernoulli trial, where the probability of success P_i is drawn at random from a beta distribution with parameters determined by $\sum_j X_{i,j}$. Formally,

$$P_i|\alpha_i, \beta_i \sim \text{beta}\left(\alpha_i = \max\left\{\varepsilon, \sum_j X_{i,j}\right\}, \beta_i = \max\{\varepsilon, n_i - \alpha_i\}\right)$$

We model the likelihood of observing k successes in $n = 20$ trials using the binomial density function as:

$$k|P_i \sim \text{binom}(P_i, n = 20)$$

Finally, the predictive distribution $p(k)$, is computed using the product of the binomial likelihood and beta prior, modelled as a 'beta–binomial distribution'[29].

$$p(k) = E[k|P_i] = \int_0^1 k|P_i \,.\, P_i|\alpha_i, \beta_i \, \mathrm{d}P_i = \binom{n}{k} \frac{B(k+\alpha_i, n-k+\beta_i)}{B(\alpha_i, \beta_i)}$$

We model the probability of sample i being ecDNA positive with the random variable Y_i so that:

$$Y_i = 1 - (k = 1|P_i) - (k = 0|P_i)$$

The expected value of Y_i is:

$$E(Y_i) = 1 - p(k = 1) - p(k = 0)$$

Let T be the set of samples belonging to a certain sample type t, for example, immortalized samples.

We define

$$Y_T = \frac{\sum_{i \in T} Y_i}{|T|}$$

We estimate the frequency of samples under sample t containing ecDNA (bar heights on Fig. 2c, d) as

$$E[Y_T] = \frac{\sum_{i \in T} E[Y_i]}{|T|}$$

and error bar heights (Fig. 2c, d) as:

$$\text{s.d.}(Y_T) = \frac{\left(\sum_{i \in T} \text{var}[Y_i]\right)^{\frac{1}{2}}}{|T|}$$

assuming independence among samples $i \in T$. For any α_i or $\beta_i = 0$, we assign them a sufficiently small ε. For more detail, please see Supplementary Information 1.

Comparison of ecDNA presence between different sample types. We construct binary ecDNA-presence distributions, based on the ecDNA counts, such that an image with ≥ 2 ecDNA is represented as a 1, and 0 otherwise. In order to compare the ecDNA presence between the different samples, we use a one-sided Wilcoxon rank-sum test using the binary ecDNA-presence distributions, where the null hypothesis assumes the mean ranks of the compared sample types are equal.

ECdetect: software for detection of extrachromosomal DNA from DAPI staining metaphase images. The software applies an initial coarse adaptive thresholding[30,31] on the DAPI images to detect the major components in the image with a window size of 150 × 150 pixels, and $T = 10\%$. Components over 3,000 pixels and 80% of solidity are masked, and small components discarded. Weakly connected components of the remaining binary image are computed to find the separate chromosomal regions. Connected components over a cumulative pixel count of 5,000 are considered as candidate search regions, and their convex hull with a dilation of 100 pixels are added into the ecDNA search region. Following the manual masking and verification of the ecDNA search region, a second finer adaptive thresholding with a window size of 20 × 20 pixels and $T = 7\%$ is performed. Components that are greater than 75 pixels are designated as non-ecDNA structures and their 15-pixel neighbourhood is removed from the ecDNA search region. Any component detected with a size less than or equal to 75 pixels and greater than or equal to 3 pixels inside the search region is detected as ecDNA. For more detail, please see Supplementary Information 2.

Bioinformatic datasets. We sequenced 117 tumour samples including 63 cell lines, 19 neurospheres and 35 cancer tissues with coverage ranging from 0.6× to 3.89× and an additional 8 normal tissues as controls. See Extended Data Fig. 4 for the coverage distribution across samples. We mapped the sequencing reads from each sample to the hg19 (GRCh37) human reference genome[32] from the UCSC genome browser[33] using BWA software version 0.7.9a (ref. 34). We inferred an initial set of copy-number variants (CNVs) from these mapped sequence samples using the ReadDepth CNV software[35] version 0.9.8.4 with parameters FDR = 0.05 and overDispersion = 1.

We downloaded CNV calls for 11,079 paired tumour–normal samples covering 33 different tumour types from TCGA. We applied similar filtering criteria to ReadDepth output and TCGA calls to eliminate false copy number amplification calls from repetitive genomic regions and hotspots for mapping artefacts.

We used the filtered set of CNV calls from ReadDepth as input probes for AmpliconArchitect which revealed the final set of amplified intervals and the architectures of the amplicons. See Supplementary Information 3 for more details.

Reconstruction using AmpliconArchitect. We developed a novel tool AmpliconArchitect, to automatically identify connected amplified genomic regions and reconstruct plausible amplicon architectures. For each sample, AmpliconArchitect takes as input an initial list of amplified intervals and whole-genome sequencing paired-end reads aligned to the human reference. It implements the following steps to reconstruct the one or more architectures for each amplicon present in the sample: (1) use discordant read-pair alignments and coverage information to iteratively visit and extend connected genomic regions with high copy numbers; (2) for each set of connected amplified regions, segment the regions based on depth of coverage using a mean-shift segmentation to detect copy-number changes and discordant read-pair clusters to identify genomic breaks; (3) construct a breakpoint graph connecting segments using discordant read-pair clusters; (4) compute a maximum-likelihood network to estimate copy counts of genomic segments; and (5) report paths and cycles in the graph that identify the dominant linear and circular structures of the amplicon (see also Supplementary Information 3).

Comparison of CNV gains between the sequencing sample set and TCGA. We compared our sample set against TCGA samples to test the assumption that the genomic intervals amplified in our sample set are broadly representative of a pan-cancer dataset, by comparing against TCGA samples. Here, we deal with an abstract notation to represent different datasets and describe a generic procedure to compare amplified regions. Consider a set of K samples. For any $k \in [1,..., K]$, let S_k denote the set of amplified intervals in sample k.

Let c be the cancer subtype for sample k. We compare S_k against TCGA samples with subtype c. Let T denote the set of all genomic regions which are amplified in at least 1% of TCGA samples of subtype c. For each interval $t \in T$, let f_t denote its frequency in TCGA samples of subtype c. We define a match score

$$d_k = \sum_{t \in S_{k,T}} f_t \quad S_{k,T} = \{t \in T \text{ s.t. } t \text{ overlaps an interval in } S_k\}$$

The cumulative match score for all samples is defined as:

$$D = \sum_{t \le k \le K} d_k$$

To compute the significance of statistic D, we do a permutation test. We generate N random permutations of the TCGA intervals for subtype c and estimate the distribution of match scores of our sample set against the random permutations. We choose a random assignment of locations of all intervals in T, while retaining their frequencies. For the jth permuted set T_j, we computed the cumulative match score D_j relative to our sample set. Thus the significance of overlap between amplified intervals in our sample set and the TCGA set is estimated by the fraction of random permutations with $D_j > D$. Computing 1 million random permutations generated exactly one permutation breaching the TCGA score D, implying a $P \le 10^{-6}$.

Oncogene enrichment. We compared the rank correlation of the most frequent oncogenes in our sample set with the top oncogenes as reported by TCGA pan-cancer analysis in ref. 20. We identified 14 oncogenes occurring in 2 or more samples of our sample set and compared these to the top 10 oncogenes from the TCGA pan-cancer analysis. We found that 7 out of the top 10 oncogenes were represented in our list of 14 oncogenes. Considering 490 oncogenes in the COSMIC database, the significance of observing 7 or more oncogenes in common in the two datasets is given by the hypergeometric probability

$$P = \sum_{i=7}^{10} \frac{\binom{480}{14-i}\binom{10}{i}}{\binom{490}{14}} = 3.07 \times 10^{-10}$$

Amplicon structure similarity. We found high similarity between amplicon structures of biological replicates (for example, Extended Data Fig. 8). We estimate the probability of common origin between two samples by measuring the pairwise similarity between amplicon structures. In reconstructing the structures (Supplementary Information 3), we identify a set of locations representing change in copy number and we use the locations of change in copy number to estimate the similarity in amplicon structures.

Let L be the total length of amplified intervals. These intervals are binned into windows of size r, resulting in $N_b = \frac{L}{r}$ bins. We use a segmentation algorithm that determines if there is a change in copy number in any bin, within a resolution of $r = 10{,}000$ bp (see meanshift in coverage: Supplementary Information 3.2.). Note that this is an overestimate, because with split-reads and high-density sequencing data, we can often get the resolution down to a few base pairs. Let S_1 and S_2 represent the set of bins with copy-number changes in the two samples, respectively. S_1 and S_2 are selected from a candidate set of locations N_b. Under the null hypothesis that S_2 is random with respect to S_1, we expect $I = S_1 \cap S_2$ to be small. Let $m = \min\{|S1|, |S2|\}$, and $M = \max\{|S_1|, |S_2|\}$. A P value is computed as follows:

$$P = \sum_{i=|I|}^{m} \frac{\binom{N_b - m}{M - i}\binom{m}{i}}{\binom{N_b}{M}}$$

When looking at GBM39 replicates (Extended Data Fig. 8), we find that all replicates displaying *EGFR* ecDNA are similar to each other. Comparing replicates in row 1 and row 2 among $|N_b| = 129$ bins (1.29 Mb), $|S1| = 5$ corresponding to row 1 (ecDNA sample), $|S2| = 6$ corresponding to row 2 (ecDNA sample) and intersection set size $|I| = 5$, we compute that the P value for observing such structural similarity by random chance is 2.18×10^{-8}, which is the highest P value among all ecDNA replicate pairs. In addition, we compare the replicates containing *EGFR* in ecDNA with the culture containing *EGFR* in HSR. Among $|N_b| = 129$ bins, $|S1| = 6$ corresponding to row 2 (ecDNA), $|S2| = 4$ corresponding to row 4 (HSR), the intersection set has size $|I| = 4$ intervals giving a P value of 1.98×10^{-5}, which gives the highest P value among the 3 ecDNA replicates compared to the HSR culture, suggesting a common origin.

A branching process model for oncogene amplification. Consider an initial population of N_0 cells, of which N_a cells contain a single extra copy of an oncogene. We model the population using a discrete generation Galton–Watson branching process[23]. In this simplified model, each cell in the current generation containing k amplicons (amplifying an oncogene) either replicates with probability b_k to create the next generation, or dies with probability $1 - b_k$ to create the next generation. We set the selective advantage

$$\frac{b_k}{1 - b_k} = \begin{cases} 1 + s f_m(k) & 0 \le k < M_a \\ 0 & \text{otherwise} \end{cases}$$

In other words, cells with k copies of the amplicon stop dividing after reaching a limit of M_a amplicons. Otherwise, they have a selective advantage for $0 < k \le M_a$, where the strength of selection is described by $f_m(k)$, as follows:

$$f_m(k) = \begin{cases} \dfrac{k}{M_s} & (0 \le k \le M_s) \\ \dfrac{1}{1 + e^{-\alpha(k-m)}} & (M_s < k < M_a) \end{cases}$$

Here, s denotes the selection coefficient, and parameters m and α are the 'mid-point', and 'steepness' parameters of the logistic function, respectively. Initially, $f_m(k)$ grows linearly, reaching a peak value of $f_m(k) = 1$ for $k = M_s$. As the viability of cells with large number of amplicons is limited by available nutrition[36], $f_m(k)$ decreases logistically in value for $k > M_s$ reaching $f_m(k) \to 0$ for $k \ge M_a$. We model the decrease by a sigmoid function with a single mid-point parameter m so that $f_m(m) = 0.5$. The 'steepness' parameter α is automatically adjusted to ensure that $\max\{1 - f_m(M_s), f_m(M_a)\} \to 0$.

The copy-number change is affected by different mechanisms for extrachromosomal (ecDNA) and intrachromosomal (HSR) models. In the ecDNA model, the available k amplicons are on ecDNA elements which replicate and segregate independently. We assume complete replication of ecDNA elements so that there are $2k$ copies which are partitioned into the two daughter cells via independent segregation. Formally, the daughter cells end up with k_1 and k_2 amplicons respectively, where

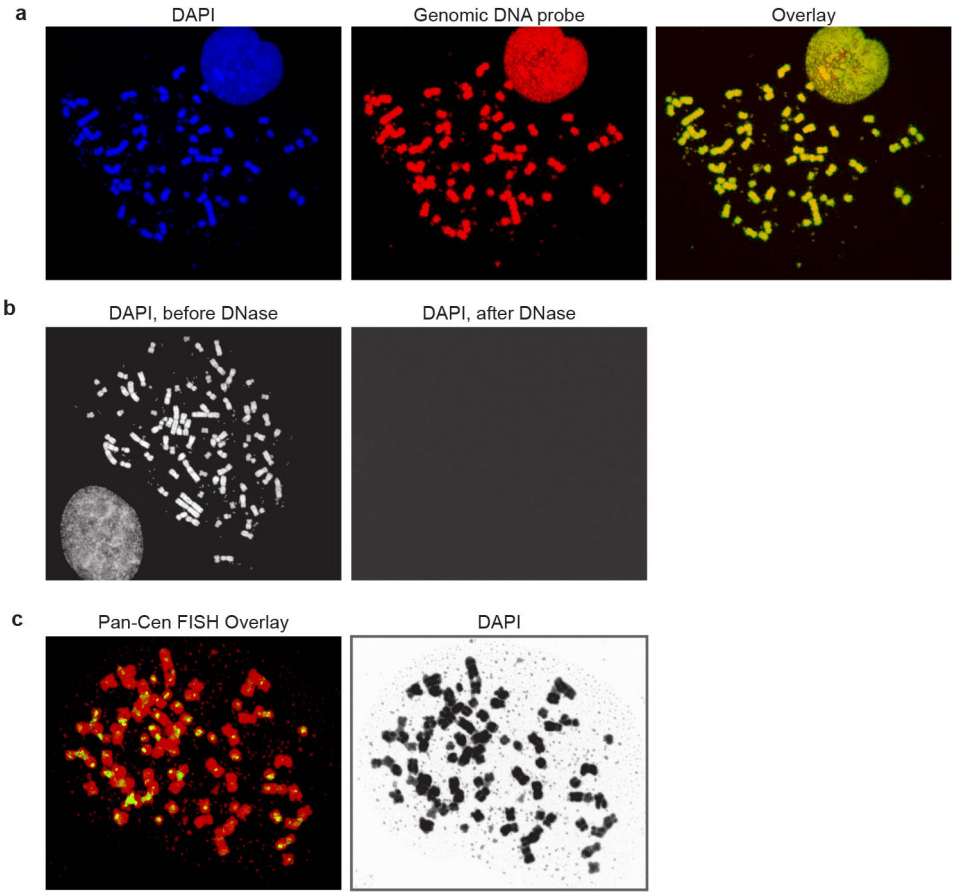
$$k_1 \sim B\left(2k, \frac{1}{2}\right)$$
$$k_2 = 2k - k_1$$

By contrast, in the intrachromosomal model, the change in copy number happens via mitotic recombination, and the daughter cell of a cell with k amplicons will acquire either $k+1$ amplicons or $k-1$ amplicons, each with probability P_d. With probability $1 - 2P_d$, the daughter cell retains k amplicons. See Supplementary Information 4 for more details.

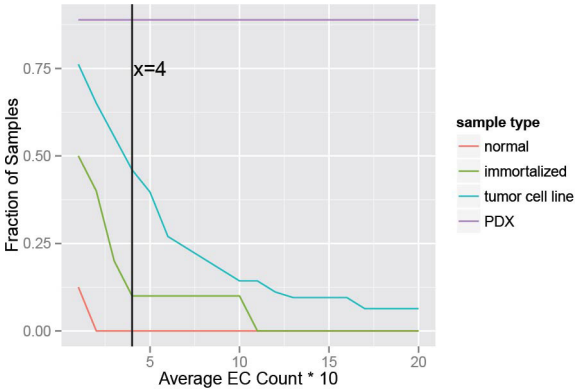
Code availability. AmpliconArchitect is available for use online at: https://github.com/virajbdeshpande/AmpliconArchitect. ECdetect will be available upon request.

Data availability. Whole-genome sequencing data are deposited in the NCBI Sequence Read Archive (SRA) under Bioproject (accession number: PRJNA338012). DAPI and FISH metaphase images are available for download on figshare at https://figshare.com/s/ab6a214738aa43833391.

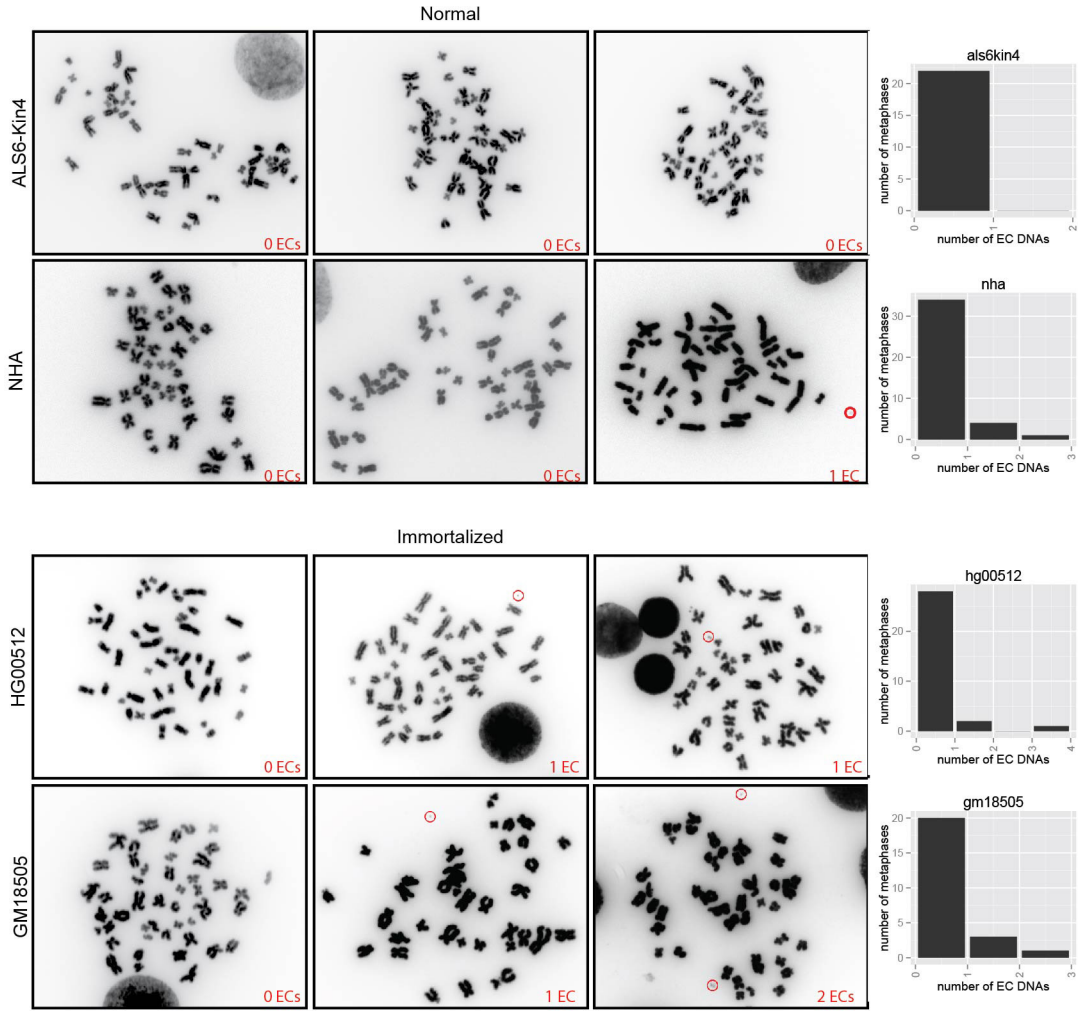
29. Lee, P. M. *Bayesian statistics: an introduction.* 4th edn (John Wiley & Sons, 2012).
30. Motl, J. *Bradley local image thresholding*. https://www.mathworks.com/matlabcentral/fileexchange/40854 (2015).
31. Bradley, D. & Roth, G. Adaptive thresholding using the integral image. *J. Graphics Tools* **12,** 13–21 (2007).
32. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
33. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12,** 996–1006 (2002).
34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
35. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6,** e16327 (2011).
36. Pavlova, N. N. & Thompson, C. B. The emerging hallmarks of cancer metabolism. *Cell Metab.* **23,** 27–47 (2016).



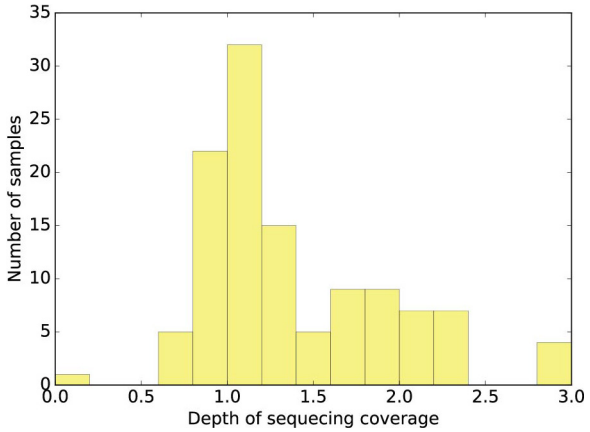
Extended Data Figure 1 | Full metaphase spreads corresponding to the partial metaphase spreads shown in Fig. 1a, Images corresponding to Fig. 1b. **b**, Images corresponding to Fig. 1c. **c**, Images corresponding to Fig. 1d.



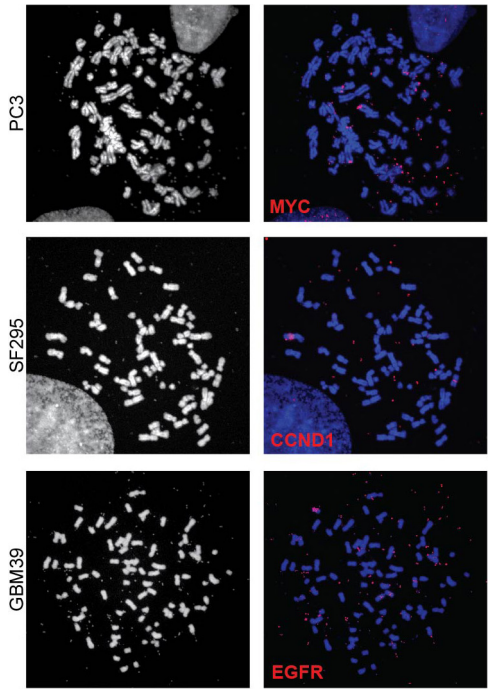
Extended Data Figure 2 | Alternative analysis of ecDNA presence according to varying criteria, stratified by sample type. Samples with a minimum number of ecDNA elements per 10 cells in metaphase in average shown in *x* axis are classified ecDNA positive, and their fraction is displayed on the *y* axis. The vertical line at $x = 4$ shows that for a minimum of 4 ecDNA elements per 10 cells in metaphase on average, 0% of normal, 10% of immortalized, 46% of tumour cell line and 89% of PDX samples are classified as ecDNA positive.



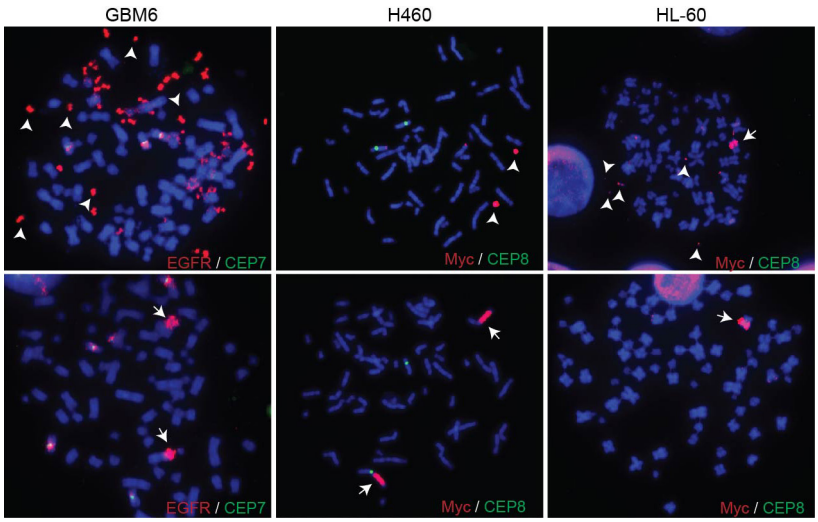
Extended Data Figure 3 | ecDNA counts in normal and immortalized cells.



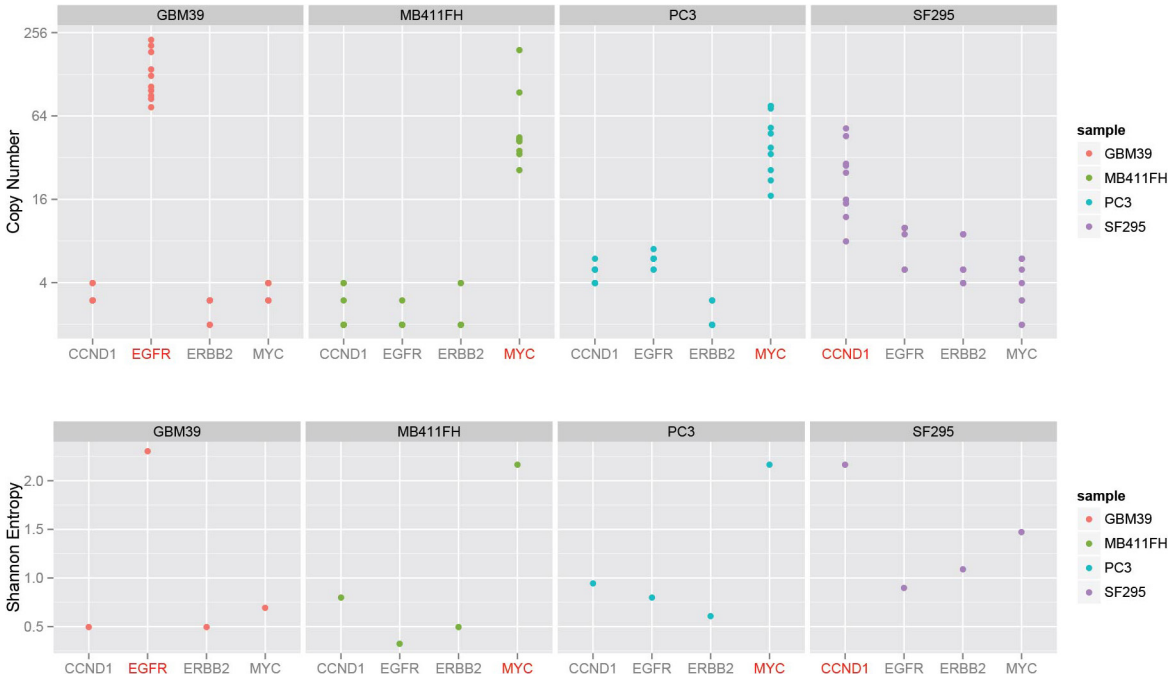
Extended Data Figure 4 | Histogram of depth of coverage for next-generation sequencing of tumour samples. We sequenced 117 tumour samples including 63 cell lines, 19 neurospheres (PDX) and 35 cancer tissues with coverage ranging from 0.6× to 3.89× (excluding one sample with 0.06× coverage) with median coverage of 1.19×.



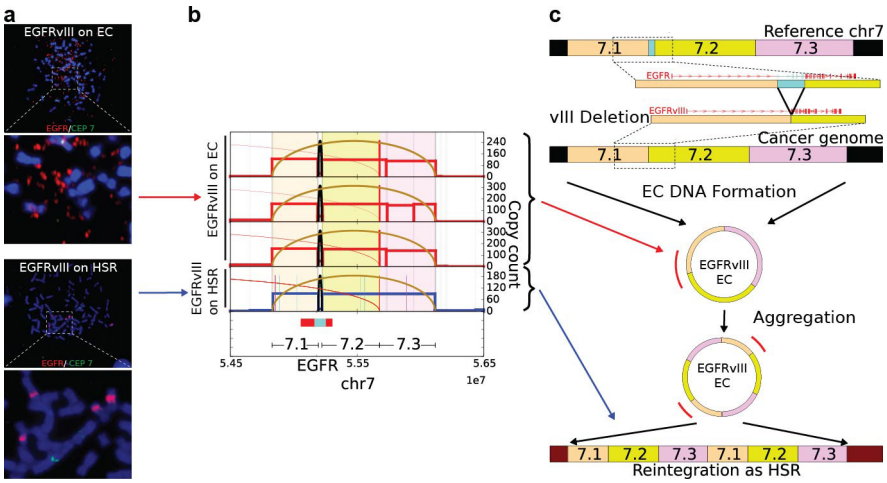
Extended Data Figure 5 | Full metaphase spreads corresponding to the partial metaphase spreads shown in Fig. 3c.



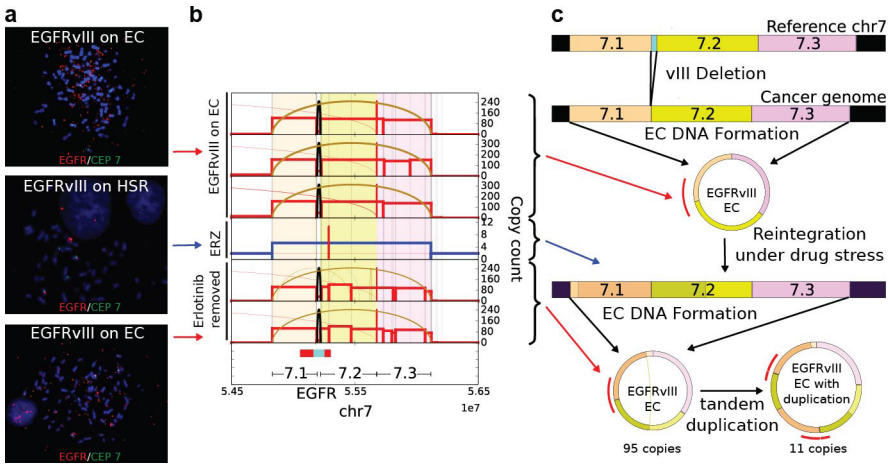
Extended Data Figure 6 | FISH images displaying both ecDNA elements and HSRs in cells from the same sample.



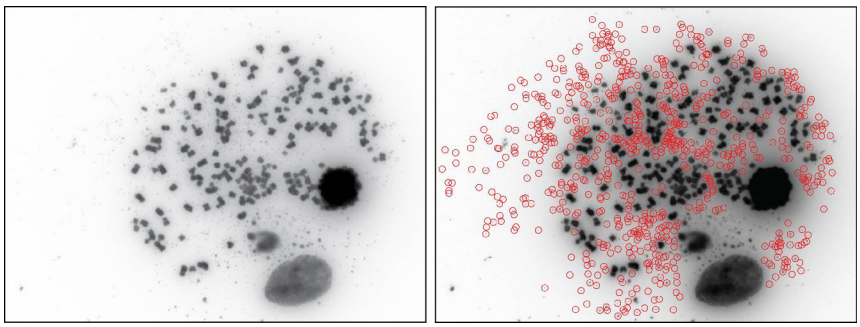
Extended Data Figure 7 | Copy-number amplification and diversity due to ecDNA. To test how much of the copy-number amplification and diversity could be attributed to ecDNA, we chose FISH probes that bind to four of the most commonly amplified oncogenes in our sample set, *EGFR*, *MYC*, *CCND1* or *ERBB2*, and quantified the cell-to-cell variability in their DNA copy number in metaphase spreads, from four tumour cell lines: GBM39, MB411FH, SF295 and PC3 cancer cells. For each cell line, only the target oncogene marked in red is known to be amplified on ecDNA (*EGFR* in GBM39; *MYC* in MB411FH and PC3, and *CCND1* in SF295). The other 3 genes reside on chromosomal loci. The target oncogene shows consistently higher copy numbers (top) and diversity (bottom).



Extended Data Figure 8 | Fine structure analysis of EGFRvIII amplification in extrachromosomal or chromosomal DNA in GBM39 cells. a, FISH images showed the *EGFR* gene on ecDNA (top) and HSRs (bottom) in different passages of the GBM39 cell line. Analysis of the HSR FISH images shows evidence of multiple integration sites on different chromosomes. **b**, Next-generation sequencing of DNA from 4 independent cultures of GBM39 was used to analyse the fine structure of amplifications (Supplementary Information 4.3). In 3 biological replicates (rows 1–3) of these cultures, EGFRvIII was exclusively on ecDNA, whereas one of the later passage cultures (row 4) was found to contain EGFRvIII entirely on HSRs, with no detectable ecDNA. The DNA derived from different ecDNA cultures shows identical structure with some heterogeneity ($P < 2.18 \times 10^{-8}$ for all pairs), suggesting common origin. However, DNA derived from HSRs show a conserved structure that is identical to the ecDNA structure ($P < 1.98 \times 10^{-5}$, Supplementary Information 2.4), possibly with tandem duplications. **c**, A possible progression of normal genome to cancer genome with EGFRvIII ecDNA and amplification to a copy count of around 100 copies. The EGFRvIII ecDNA elements possibly aggregate into tandem duplications and reintegrate into multiple chromosomes as HSRs, so that 5–6 HSRs accommodate around 100 copies of EGFRvIII.



Extended Data Figure 9 | Fine structure analysis of EGFRvIII amplification in extrachromosomal or chromosomal DNA in naive GBM39 cells and in response to erlotinib treatment and drug withdrawal. a, FISH images of naive GBM39 cells, in response to erlotinib treatment (ERZ) and drug withdrawal (after ERZ is removed) displayed ecDNA amplification, HSR amplification and ecDNA amplification, respectively (top to bottom). **b**, Next-generation sequencing of DNA from 6 independent cultures of GBM39 was used to analyse the fine structure of amplifications (Supplementary Information 4.3). Average copy numbers of amplified intervals as determined from sequencing analysis in naive samples: 110–150 (biological replicates in rows 1–3), the ERZ sample: 5.4 (row 4) and after drug withdrawal: 100–105 (biological replicates in rows 5–6). All three categories show similar fine structure indicating common origin (Methods). Drug-withdrawal replicates show additional rearrangements and heterogeneity as compared to naive samples. **c**, Cytogenetic and sequencing progression suggests that the EGFRvIII ecDNA in naive cells get reintegrated into HSRs after drug application and that the copies in the HSRs break off from the chromosomes again to form ecDNA with a copy count similar to naive cells. Drug-withdrawal samples also show additional heterogeneity in structure.



Extended Data Figure 10 | A GBM cell in metaphase with large ecDNA counts (>600), as determined by manual counting and ECdetect.