Evolution of oncogene amplification across 86,000 cancer cell genomes

Authors

Jake June-Koo Lee[1,2,3]*†, Sohrab Salehi[1,2,4]*, Matthew A. Myers[1], Marc J. Williams[1], Melissa A. Yao[5,6], Duaa H. Al-Rawi[1,2], Jin Lee[2], Eric G. Sun[5,7,8], Kerstin Thol[1], Seongmin Choi[1], Eliyahu Havasov[1], Asher Preska Steinberg[1], Michelle Wu[1], Nicole Rusk[1], Caitlin Timmons[1,8], Cheryl Zi Jin Phua[1,8], Stephen Martis[1], Neeman Mohibullah[9], Parvathy Manoj[2], Esther Redin[2], Álvaro Quintanal-Villalonga[2], Pedram Razavi[2], Samuel Aparicio[10,11], Natasha Rekhtman[12], Viviane Tabar[5,13], Mark M. Awad[2,3], Helena A. Yu[2,3], Kenny Kwok Hei Yu[5,13], Andrea Ventura[5], Andrew McPherson[1], Charles M. Rudin[2,3], Sohrab P. Shah[1,3]†

Affiliations

1. Halvorsen Center for Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
2. Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
3. Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medicine, New York, NY, USA.
4. Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA.
5. Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
6. Louis V. Gerstner Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
7. Tri-Institutional MD-PhD Program, Weill Cornell Medicine, Rockefeller University, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
8. Tri-Institutional PhD Program in Computational Biology & Medicine, Weill Cornell Medicine, New York, NY, USA.
9. Integrated Genomics Operation, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
10. Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, BC, Canada.
11. Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada.
12. Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
13. Department of Neurosurgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

* These authors contributed equally to this work.

† Corresponding authors: S.P.S. (shahs3@mskcc.org), J.J.L. (leej39@mskcc.org).

Abstract

High-level copy-number (CN) amplification (HLAMP) is a major mechanism of oncogene activation in human cancer. Despite progress in therapeutically targeting amplified oncogenes, the processes underlying amplicon evolution remain incompletely understood, leaving critical knowledge gaps in their etiology and mechanisms of therapeutic response. To address this, we analyzed the evolutionary trajectories of HLAMPs using single-cell whole-genome sequencing data from 86,239 cancer cells across 93 patients and 9 experimental systems. We found that cell-to-cell CN variability provides a quantifiable readout of HLAMP mechanism, clearly distinguishing extrachromosomal circular DNA (ecDNA) from intrachromosomal amplification (ICamp) through characteristic CN distributions that reflect distinct modes of segregation and correspond to clonal architecture. Notably, ICamp events frequently showed multiple amplitude peaks specific to subclones, indicating punctuated shifts in oncogene dosage through numeric or structural modulatory mechanisms with transcriptional impact. In contrast, ecDNAs exhibited broad, continuous CN distribution with extreme high-copy outliers, consistent with asymmetric segregation. The CN and structural diversity of ecDNA regions enabled systematic deconvolution of ecDNA subspecies and estimation of their per-cell abundance, revealing the history of ecDNA-mediated oncogenesis at single-nucleotide resolution. We observed ecDNA diversification through internal rearrangements across cases and, notably, convergent evolution in glioblastoma cases marked by multiple, recurrent acquisition of *EGFR*-targeting ecDNAs. Finally, single-cell genome-based identification of ecDNAs showed substantial discrepancy with bulk genome graph-based predictions and reliably distinguished actively maintained ecDNAs from historical genomic footprints after chromosomal re-integration. These findings reveal marked tissue-type specificity of ecDNAs, suggesting that ecDNA-mediated oncogenesis may depend on a permissive tissue context.

Introduction

High-level copy-number amplification (HLAMP) is a common mechanism of oncogene activation in human cancers and often creates a targetable functional dependence. For example, anti-HER2 therapies in HER2-positive breast cancer[1] and other cancer types[2,3] have enabled effective precision cancer medicine in previously intractable tumors. In contrast to other primary oncogenic alterations, such as point mutations and gene fusions, HLAMPs often exhibit profound intratumoral heterogeneity. For example, *ERBB2* (encoding HER2) copy number can differ between primary and metastatic sites[4], across regions within a tumor[5], and among individual cells[6]. Such heterogeneity has been associated with poor clinical outcomes and therapy resistance[7], yet its mechanistic basis remains incompletely understood. One plausible explanation is amplification via extrachromosomal circular DNA (ecDNA), which lacks centromeres and therefore can

undergo unequal segregation during cell division[8]. Recent studies indicate that non-Mendelian inheritance[9], together with unique biophysical properties of ecDNA[10], can drive clonal evolution and lead to heterogeneous cancer cell populations[11,12]. By contrast, intrachromosomal amplification (ICamp) is expected to segregate symmetrically with chromosomes. How ICamps evolve and contribute to observed cellular heterogeneity remains unclear. Recently, therapeutic approaches exploiting selective vulnerabilities of ecDNA-containing cancer cells have been proposed[13–15], underscoring the clinical importance of accurately distinguishing HLAMP mechanisms and characterizing their evolutionary dynamics.

To map the mechanisms underlying HLAMP etiology, evolution and their intratumoral heterogeneity in clinical tumor samples, we analyzed a large-scale single-cell whole-genome sequencing (scWGS) dataset consisting of >86,000 single cells genomes derived from >100 samples across a range of solid cancer types and experimental models. This enabled both direct quantification of cell-to-cell copy number (CN) variability[16,17] and identification of structural variant (SV)-resolved segment boundaries at nucleotide resolution at HLAMP loci. From these data, we developed a probabilistic mixture model, `eicicle`, informed by evolutionary simulations to distinguish between ICamp and ecDNA based on single-cell CN distributions. We found that ICamps diversify at a clone level pervasively through two mechanisms: numeric modulation whereby the copy number architecture is maintained, but arm-level or chromosomal-level aneuploidies increase or decrease copy number in a clone-specific manner; and structural modulation whereby SV-driven mutational processes drive subclonal diversification. In ecDNAs we observed cases harboring multiple ecDNA species, and developed `ECADeMix`, a mixed-integer linear programming framework, to decompose ecDNA heteroplasmy at single-cell resolution through simultaneous inference of ecDNA structures and cellular prevalence reflective of evolutionary histories. Finally, our scWGS-based approach underscores limitations of bulk genome graph-based ecDNA predictions and reveals notable tissue-type specificity[18] of ecDNAs, informing patient selection for emerging ecDNA-directed therapeutics.

Single-cell copy number variance distinguishes ecDNA and ICamp

We analyzed direct library preparation (DLP+)-based scWGS data from 102 individuals (patients and experimental models), including 71 published[17,19–21] and 31 newly generated cases (**Fig. 1a, top** and **Supplementary Table 1**). These included clinical tumor samples and patient-derived xenografts (PDXs) from 93 patients across four cancer types: high-grade serous ovarian cancer (HGSOC; n=61), invasive breast cancer including hormone receptor-negative (HR− BC; n=10; one with *ERBB2* amplification) and -positive (HR+ BC; n=5), lung cancer including small-cell (SCLC; n=8) and non-small cell (NSCLC; n=5; all with *EGFR* mutation), and glioblastoma multiforme (GBM; n=4). Although most clinical tumors were obtained prior to systemic therapy,

the cohort included several pretreated tumors, including 4 HR+ BCs and 2 EGFR inhibitor-resistant, *EGFR*-mutated NSCLCs (**Supplementary Table 2**). The cohort also included many PDX samples, which were often collected after chemotherapy exposure. In addition, we studied 9 experimental models, including cell lines (n=8) and a mouse model of liver cancer harboring engineered *Mdm2*-containing ecDNA[22] (n=1). A total of 311 scWGS libraries generated from these cases (**Supplementary Table 3**) were assessed for quality and jointly processed with matched normal bulk WGS data for somatic variant calling, as described previously[20]. After removing normal cells and low-quality cells, the dataset consisted of 86,239 high-quality cancer cells (**Fig. 1b**). To systematically analyze the generative and evolutionary mechanisms of HLAMP[23], we integrated somatic SV breakpoints detected at the pseudobulk level with genome-wide CN bins to generate SV-aware single-cell CN profiles (**Methods**). Using pairwise CN distance between single cells, we identified subclones and reconstructed their evolutionary relationships (**Methods**). Across the 102 cases, we identified 507 HLAMP regions with i) CN at least 3 times the chromosome arm-level baseline in one or more subclones and ii) a pseudobulk CN of at least 6. These HLAMP regions typically included *bona fide* oncogenes known to be activated by HLAMP, such as *MYC*, *CCNE1*, *EGFR*, and *MDM2* (**Fig. 1c**), in addition to loci devoid of known oncogenic drivers, often amplified in conjunction with other oncogene-containing regions. For each HLAMP region (comprising one or more contiguous SV-aware segments), we evaluated CN distributions across cells, determined generative mechanisms from SV features at their boundaries, and analyzed clone-specific features to delineate mechanisms driving subclonal diversification (**Fig. 1a, bottom**).

First, we performed *in silico* simulation of single-cell CN distributions accounting for the mode of inheritance of oncogene amplifications and evolutionary constraints (**Methods**)[9,12,24,25]. These simulations showed that conditioned on symmetric cell division, as expected for ICamps, CN distributions consist of prominent, narrow peaks reflecting cell populations that share one or a discrete set of CN values (**Fig. 1d)**. By contrast, conditioned on asymmetric cell divisions, as expected for ecDNA[12], broad, heavy-tail distribution with extreme outliers were observed, reflecting CN variance across the cell population. Modulating key parameters further simulated fitness consequences on ecDNA CN distributions: stronger positive selection (e.g., higher potency of oncogene) shifted the modal CN to the right, while greater cell-death penalties (e.g., DNA damage responses at extreme CN) shifted it to the left (**Extended Data Fig. 1** and **Methods**). Consistent with these simulations, we observed two distinct and quantitatively distinguishable classes of single-cell CN distributions across HLAMP regions in our cohort. For example, in the *EGFR*-mutated lung cancer case EL011, the *EGFR* allele harboring the E746-A750 microdeletion was highly amplified and exhibited a narrow, high-CN peak from 591 cancer cells, consistent with ICamp (**Fig. 1e, left** and **Extended Data Fig. 2a**). In contrast, the surgically resected glioblastoma GBM0721 showed a broad, heavy-tail distribution of wild-type *EGFR* CN across 522 cancer cells,

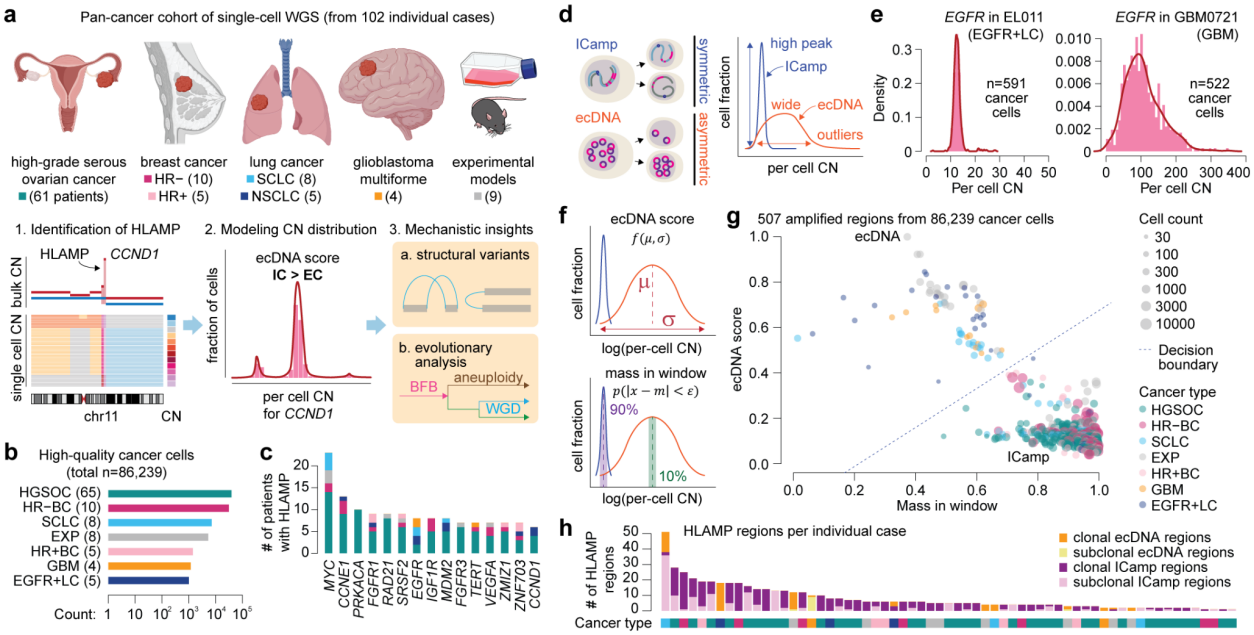


Fig. 1. Cell-to-cell copy-number variance in scWGS differentiates ecDNA-mediated amplifications from intrachromosomal amplifications. a, Cancer types and case counts (top) and schematic illustration of study design (bottom). Organ and model illustrations are from BioRender. **b**, Number of high-quality cancer cells analyzed (x-axis, log-scale) by tumor type (y-axis). **c**, Frequently amplified oncogenes and their case counts. **d**, Inheritance patterns for intrachromosomal amplification (ICamp) versus extrachromosomal circular DNA (ecDNA; **left**) and expected CN distributions across cells (right). **e**, Representative examples of ICamp and ecDNA, from a patient with *EGFR*-mutated lung adenocarcinoma (left) and glioblastoma (right). **f**, Conceptual illustration of two metrics used to classify HLAMP events. **g**, Classification of 507 HLAMP regions from 86,239 cancer cells based on single-cell CN distribution. **h**, Fraction of clonal and subclonal HLAMP regions by mechanism, across cases.

consistent with ecDNA (**Fig. 1e, right** and **Extended Data Fig. 2b**). Motivated by these patterns, we developed a machine learning-based classifier, `eicicle`, to distinguish ecDNA from ICamp (**Fig. 1f**). `Eicicle` includes two metrics: *ecDNA score* which captures distribution width, and *mass-in-window* which quantifies the fraction of cells near the modal CN (high in ICamp, low in ecDNA). These metrics are derived from a probabilistic mixture model that conditionally models the two distinctive distributions (**Methods**). The 507 HLAMP regions by these metrics were bimodally distributed (**Fig. 1g**). We performed a discriminant analysis to infer a decision boundary, designating 15 HLAMP regions with high ecDNA scores and low mass-in-window (left upper corner) as representative ecDNA and 6 regions with low ecDNA scores and high mass-in-window (right lower corner) as representative ICamp (**Extended Data Fig. 3**). This boundary classified 72 HLAMP regions from 11 cases as ecDNA and 435 regions from 55 cases as ICamp. We then orthogonally validated classifier predictions across all available experimental models (5 ecDNA-positive and 3 ecDNA-negative by `eicicle`; **Extended Data Fig. 4a-c**) using metaphase

170 DNA FISH. In all models classified as ecDNA, we observed dispersed extrachromosomal
171 fluorescent signals for the amplified oncogenes, consistent with ecDNA. In models classified as
172 ICamp, we observed strong intrachromosomal clustered signals. Notably, HLAMP regions
173 classified as ecDNA were largely from GBM, SCLC, and *EGFR*-mutated NSCLC. Despite
174 numerous HLAMP events in triple-negative breast cancers and ovarian cancers, none exhibited a
175 characteristic single-cell CN distribution for ecDNA. Instead, they showed prominent peaks
176 typical for ICamp. We then integrated ecDNA/ICamp classification with subclonal architectures
177 for each case (**Fig. 1h**). Many ecDNA-mediated HLAMP regions were clonally amplified, with 11
178 of 72 (15%) classified as subclonal (amplified in <90% of the cells). In contrast, 143 of 435 (33%)
179 ICamp regions were subclonal, indicating substantial intratumoral heterogeneity. Together, these
180 results establish single-cell CN distributions derived from scWGS as an effective quantitative
181 readout to distinguish ecDNA from ICamp, reflecting their distinct mechanistic modes of
182 inheritance during cell division, and their impact on HLAMP evolution.

183

184 **Evolution of ICamps via structural and numeric modulators**

185 Among 435 HLAMP events classified as ICamp, 270 (62%) exhibited multiple CN peaks (**Fig.**
186 **2a**), indicating intratumoral heterogeneity (exemplified by case OV-044; **Fig. 2b**). Distinct
187 HLAMP CN peaks generally aligned with phylogenetic clades defined by genome-wide CN
188 profiles. In cases with multiple peaks, individual clones were preferentially concentrated within a
189 single peak rather than distributed across multiple peaks ($p<1\times10^{-16}$, Stouffer-combined one-sided
190 permutation test). In aggregate, 69% of subclones were restricted to one peak, consistent with
191 clonal expansions subsequent to subclone-specific ICamp events. Reflecting this heterogeneity,
192 ICamp cases showed significantly greater CN variance than stable genomic regions used as
193 controls (*TP53* and *RB1*; mean standard deviation 3.32 vs. 0.575; $p=1.97\times10^{-72}$ by one-sided two-
194 sample *t*-test), although it was significantly less than ecDNA (mean standard deviation: ecDNA
195 43.1, ICamp 3.32; $p=8.68\times10^{-20}$ by one-sided two-sample *t*-test; **Fig. 2c**). To investigate the source
196 of this variability, we classified ICamp events by their generative mechanisms using boundary SV
197 features (**Extended Data Fig. 5a-f**; **Methods**) and asked whether variability differed by initiating
198 mechanism[23,26]. Chromothripsis- and breakage-fusion-bridge (BFB) cycle-mediated ICamps
199 showed greater variance than other groups combined, but the absolute difference was modest
200 (mean standard deviation 3.65 vs. 3.14; one-sided two-sample *t*-test, p=0.0253; **Fig. 2c**), indicating
201 that the initiating mutational processes may have only a small impact on CN variance. To identify
202 processes modulating ICamp CN after their initiation, we compared CN profiles between
203 subclones (**Fig. 2d-e**). We identified pairs of subclones with nearly identical chromosomal or arm-
204 level CN profiles that differed primarily in amplitude, consistent with numeric modulation (4,839
205 of 8,486 region-subclone pairs, 57%; **Fig. 2f**); these patterns occurred

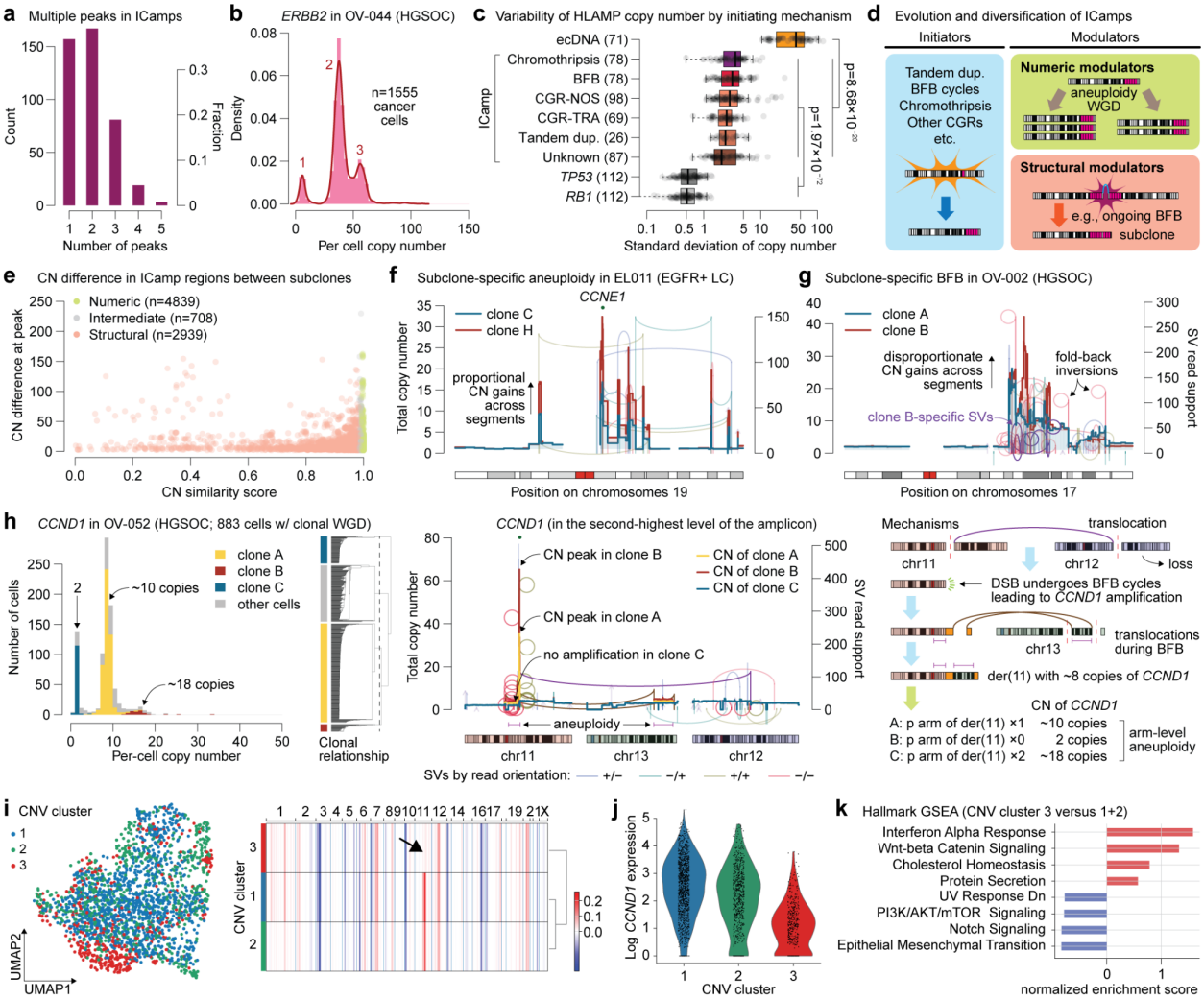


Fig. 2. Contrasting mechanisms of ICamp CN heterogeneity by numeric and structural modulators. a, Prevalence of multi-peak CN distribution among ICamp cases. **b**, An ICamp case exhibiting multi-peak CN distribution across cells, indicating subclones with distinct *ERBB2* CN. **c**, CN variance of HLAMP events stratified by initiating mechanism (BFB, breakage-fusion bridge; CGR-NOS, complex genomic rearrangement, not otherwise specified; CGR-TRA, complex genomic rearrangement, translocation-rich). **d**, Initiating and modulating mechanisms for ICamps. **e**, Pairwise comparisons of HLAMP-associated arm-level CN profiles between subclones. **f**, A classical example of aneuploidy-mediated doubling of *CCNE1* amplification in EL011. Compared to clone C, all the HLAMP segments exhibit proportional CN increase in clone H. **g**, A structural modulation of 17q amplification in OV-002. Clone B exhibits further amplification of two segments (one amplifying *MIR21*) within the HLAMP segment in clone A. **h**, Multiple CN peaks of *CCND1* amplification in OV-052 matching the subclonal phylogeny (left), clone-specific CNs and SVs indicative of BFB-mediated amplification and subsequent modulation by arm-level aneuploidy (middle), and schematics of the mechanism (right). **i**, UMAP embedding of 2,244 cancer cells from scRNA sequencing of OV-052 (left) and summary CN profiles of inferred clones in the malignant compartment (right). A subclone with no *CCND1* amplification (cluster 3; n=431 cells) is detected. **j**, Normalized expression level of *CCND1* in each cluster. **k**, Gene set enrichment analysis comparing cluster 3 with the other clusters. The top 8 enriched pathways are shown.

223 most frequently in the context of whole-genome doubling (WGD; 2,033/4,839, 42%; **Extended**
224 **Data Fig. 6a**), and the rest from chromosome- or arm-specific aneuploidies (**Extended Data Fig.**
225 **6b)**. In other cases, subclones showed dissimilar CN profiles, often accompanied by subclone-
226 specific SVs, indicative of structural modulation (2,939/8,486, 35%; **Fig. 2g**). These involved
227 diverse mechanisms, including subclone-specific fold-back inversion (consistent with additional
228 BFB cycles), chromothripsis, and other complex genomic rearrangements (**Extended Data Fig.**
229 **7**).

230 As an illustrative example, HGSOC case OV-052 harbored a *CCND1* ICamp with two large and
231 one small CN peaks, indicative of a heterogeneous ICamp. These peaks aligned with three
232 subclones in the CN-based phylogeny (**Fig. 2h, left**). In the major clone A, the *CCND1*
233 amplification, adjacent to an inter-chromosomal rearrangement translocating the distal portion of
234 11q to 12q, exhibited multiple fold-back inversions at its borders, consistent with BFB cycles
235 further modifying locus subsequent to the translocation. Minor subclone B exhibited arm-level
236 aneuploidy that duplicated the *CCND1* amplicon, reflected by two-fold CN across segments from
237 the centromere through the distal amplicon (**Fig. 2h, middle and right**). Notably, minor clone C
238 lacked evidence of the *CCND1* amplification and showed a subclone-specific loss of the allele
239 with the amplification. Analysis of haplotype-informed allele-specific CN was consistent with
240 arm-level aneuploidy-mediated loss of the *CCND1* ICamp (**Extended Data Fig. 6c**). Moreover,
241 single-cell RNA sequencing from adjacent tumor tissue identified a subclone with negligible
242 cyclin D1 expression and a genome-wide CN profile matching clone C, supporting the phenotypic
243 impact of this aneuploidy-mediated HLAMP loss (**Fig. 2i-j**). Notably, this subclone exhibited a
244 distinct gene expression program, with a high interferon alpha response and upregulation of
245 antigen-presentation machinery (**Fig. 2k**), features proposed to promote an immunosuppressive
246 microenvironment[27]. These results suggest that such bidirectional numeric modulation of ICamp
247 could enhance transcriptional adaptability under selection. We observed similar aneuploidy-
248 mediated elimination of ICamp events in 3 additional cases, often involving G1/S cyclin genes
249 (**Extended Data Fig. 6c**).

250 We also observed diversification of chromothripsis-mediated HLAMPs through both numeric and
251 structural mechanisms. In one representative case (OV-083), chromosome 8 underwent two rounds
252 of chromothriptic complex rearrangement, generating two subclones that were each further
253 diversified by aneuploidy (**Extended Data Fig. 8a**). In another case (OV-004), chromothripsis of
254 chromosome 8 gave rise to two clades with distinct DNA repair configurations derived from the
255 same intermediate chromosome; as in OV-083, both clades were subsequently diversified by
256 aneuploidy (**Extended Data Fig. 8b**).

257 Together, our analyses highlight punctuated evolution of ICamp through two principal routes[28]:
258 numeric modulation by aneuploidy and WGD, and structural remodeling via BFB, chromothripsis,

259 and other mutational processes. Notably, these changes occur as both gains and losses and often
260 define sizable subclones, reflective of non-negative fitness and a linkage to tumor evolution.

261

262 **ecDNA architecture resolved by cell-to-cell CN variance and genomic rearrangement**

263 We identified 72 HLAMP regions across 11 cases (6 patients and 5 experimental models) with
264 single-cell CN distributions consistent with ecDNA. Notably, genome-wide CN profiles in these
265 11 cases were relatively homogeneous across cells, with most intercellular differences confined to
266 HLAMP regions. Given the presence of multiple HLAMP segments in each ecDNA-positive case,
267 we first leveraged within-case single-cell CN correlations to infer whether segment pairs were co-
268 located in the same ecDNA species or segregated independently. In GBM0721, two ecDNA-
269 mediated HLAMP segments were identified: one encompassing *MDM4* (1q32.1) and another
270 *EGFR* (7p11.2) (**Fig. 3a, left**). They showed a modest but significant correlation, with considerable
271 variability across the CN range (**Fig. 3a, right**; Pearson correlation=0.25; $p=6.5\times10^{-8}$). In contrast,
272 in a surgically resected, heavily pretreated *EGFR*-mutated lung cancer case EL001 where we
273 identified 18 ecDNA-amplified segments, *EGFR* (7q11.2) and *CCND1* (11q13.3) CNs were tightly
274 and linearly correlated despite their different chromosomes of origin (**Fig. 3b;** Pearson
275 correlation=0.98; $p=3.4\times10^{-58}$). In GBM0721, the *MDM4* and *EGFR* amplicons likely reside on
276 separate ecDNA species, consistent with independent head-to-tail SVs encircling each segment
277 (**Fig. 3c**). The correlation between *MDM4* and *EGFR* CN may reflect co-selection for a fitness
278 advantage, supported by single-cell RNA-seq from an adjacent tumor region: cells with high
279 *MDM4* and *EGFR* expression showed high proliferative features (e.g., mitotic spindle, Ki-67),
280 whereas cells with low expression of both showed elevated reactive oxygen species signaling
281 (**Extended Data Fig. 9a-c**). By contrast, the linear *EGFR*/*CCND1* correlation in EL001 suggests
282 a chimeric ecDNA species carrying both oncogenes, supported by chromothripsis connecting
283 segments harboring *EGFR* and *CCND1* (**Fig. 3d**). Similarly, *CDK4* and *MDM2* CNs were also
284 tightly correlated (**Extended Data Fig. 10a**) with SV links, implying another ecDNA species co-
285 amplifying these oncogenes. Recognizing that the copy number of segments on the same ecDNA
286 species will co-vary across cells, we computed pairwise correlations among all HLAMP regions
287 in EL001 and performed hierarchical clustering (**Fig. 3e; Methods**). This revealed two largely
288 mutually exclusive correlation clusters, one comprising regions from chromosomes 7 and 11, and
289 another with chromosome 12 regions, consistent with two distinct ecDNA species each harboring
290 two genes.

291 Motivated by these observations, we developed `ECADeMix` (Extra-Chromosomal Amplicon De-
292 Mixing) to systematically identify the structure of ecDNA species and their heteroplasmy across
293 cells. Briefly, analogous to heteroplasmy analyses in mitochondrial DNA in single cells[29],

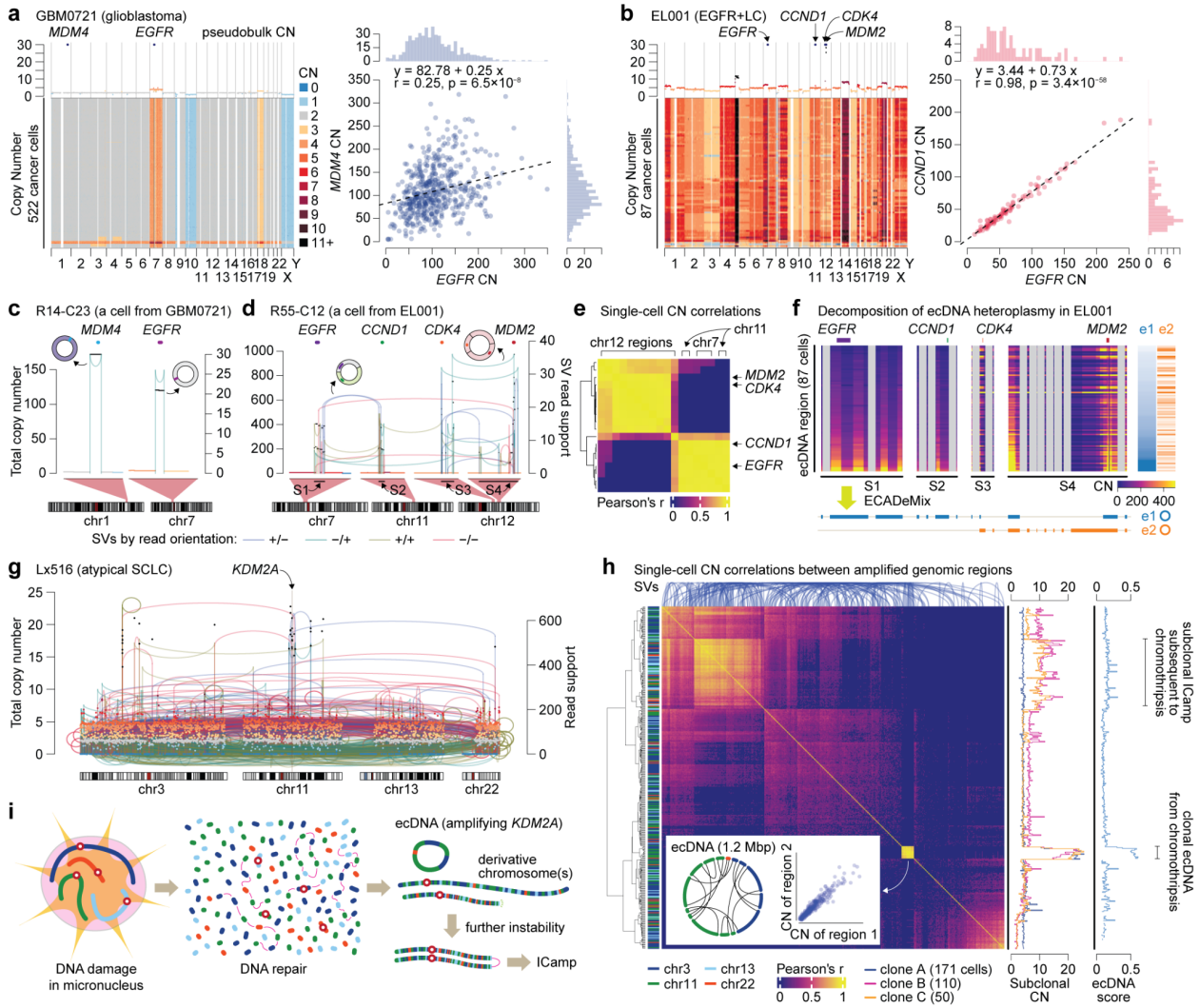


Fig. 3. Resolving mechanisms of segregation of co-amplified oncogenes in ecDNA in single cells. a-b, Genome-wide CN profiles (left) and CN correlations between key oncogenes (right) for GBM0721 and EL001. **c-d**, Single cell CN/SV profiles indicating two distinct simple ecDNAs in GBM0721 and two distinct hybrid ecDNAs in EL001. **e**, Hierarchical clustering of HLAMP regions in EL001 based on pairwise CN correlations reveals two ecDNA species co-amplifying oncogenes. **f**, Decomposition of ecDNA heteroplasmy in EL001 by `ECADeMix`. HLAMP segments used (S1-S4) are indicated in **d**. **g**, Pseudobulk CN/SV profile from Lx516 showing multi-chromosomal chromothripsis leading to HLAMP of *KDM2A*. **h**, Clustering based on pairwise CN correlation identifies a hybrid ecDNA and a subclonal ICamp event. Rearrangements between segments are illustrated as arcs on the top. Chromosomal origins in the left heatmap. Subclonal CN and ecDNA scores on the right side. Inset plots show the reconstructed ecDNA structure (left) and a representative CN correlation between two segments that comprise the ecDNA. **i**, Schematic illustration of HLAMP mechanisms in Lx516.

`ECADeMix` jointly infers integer CN profiles for each ecDNA species and their multiplicities per cell by decomposing single-cell CN profiles of ecDNA regions: here to 10-Kbp resolution

309 (**Extended Data Fig. 10-12**; **Methods**). ECADeMix identified two ecDNA species in each of
310 GBM0721 and EL001, confirming our previous conclusions and additionally revealing that in
311 EL001 the *MDM2* locus was also present on the *EGFR*/*CCND1* ecDNA (**Fig. 3f**), a feature that
312 could not be resolved from correlation analyses. Furthermore, ECADeMix identified ecDNA
313 species co-amplifying multiple oncogenes in two other cases (*MYC* and *CDX2* in COLO320-DM
314 and *NKX2-1* and *FOXA1* in EL003) as well as those in which multiple distinct clone-specific
315 species separately amplified the same oncogene (e.g., GBM0510, discussed below). Using
316 ECADeMix, we identified a total of 23 ecDNA species across 11 cases, including 6 cases with
317 multiple species (2-11 each). This included surgically resected brain metastases of atypical SCLC
318 characterized by an exceptional multi-chromosomal chromothripsis involving four chromosomes,
319 resulting in ecDNA-mediated HLAMP (**Fig. 3g**). Given the extreme complexity of this case, we
320 questioned how many ecDNA species had formed through this process. Notably, ECADeMix
321 identified only a single ecDNA species conjoining 13 regions originating from 3 chromosomes
322 (**Extended Data Fig. 10d**). Accordingly, pairwise CN correlations among all 391 amplified
323 regions in this case (152,881 pairs) revealed a distinct cluster of near-perfect correlations
324 encompassing all 13 regions that formed a full circle in structural reconstruction (**Fig. 3h)**.
325 Additional regions showed strong but more modest correlations (left upper part of the heatmap);
326 their CN profiles and ecDNA scores are consistent with subclonal ICamp events arising from
327 chromothriptic derivative chromosome(s), likely reflecting ongoing instability (**Fig. 3i**). Together,
328 these analyses illustrate that cell-to-cell CN variance and pairwise segment correlations enable
329 deconvolution of complex ecDNA architectures. Further, ECADeMix analysis of scWGS unveils
330 new evolutionary models for ecDNA through robust estimates of both presence and prevalence of
331 ecDNA species in each cell, enabling an algorithmic approach to dissect complex ecDNA-
332 mediated cancer evolution.

333

334 **ecDNA evolution at single-nucleotide resolution**

335 We next integrated SV features and amplicon species information at the single-cell level to
336 elucidate how ecDNA drives cancer evolution. The high coverage of ecDNA regions enabled
337 single cell-level, nucleotide-level analysis of SVs (median per-cell coverage 11.09X, range
338 0.9522X−48.11X). In 5 of 11 ecDNA-positive cases (Lx33, COLO320-DM, GBM39-DM, NCI-
339 H69, NCI-H524, 2765_2), we detected a dominant ecDNA species accompanied by subpopulation
340 of cells sharing lower-frequency variant forms, consistent with branched (divergent) evolution that
341 followed initial clonal expansion by the ancestral ecDNA (**Extended Data Fig. 11a-c**). For
342 example, the NCI-H69 SCLC cell line harbored a chromothripsis-derived ecDNA amplifying
343 *MYCN* (**Fig. 4a, top left**) with more than half of cells carrying a predominant parental ecDNA
344 species (e0) containing all chromothriptic HLAMP segments. Using ECADeMix, we identified

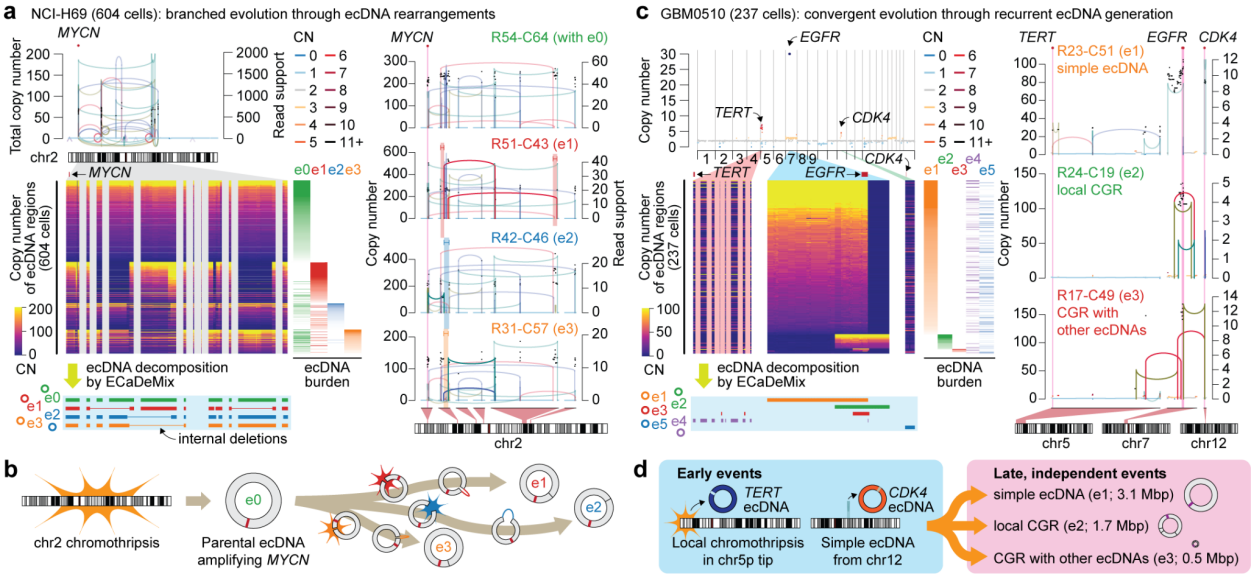


Fig. 4. Modes of ongoing ecDNA evolution. a, Chromothripsis-derived ecDNA amplifying *MYCN* in the NCI-H69 SCLC cell line (top left); single-cell CN profiles of the amplicon and ecDNA heteroplasmy delineated by `ECADeMix` (bottom left). CN/SV profiles from 4 cancer cells with distinct ecDNA subspecies (right). e0 is the likely parental ecDNA; e1-e3 are daughter ecDNAs derived from e0 via internal rearrangements. **b**, Schematic illustration of branched evolution via ecDNA rearrangement. **c-d**, Pseudobulk genome-wide CN profile of glioblastoma GBM0510, exhibiting multiple ecDNA-mediated HLAMPs (**c,** top left); single-cell CN profiles of the amplicons and decomposed ecDNA species (**c,** bottom left). *EGFR* is amplified by three distinct ecDNAs (e1, e2, and e3) in a mutually exclusive manner across cells, forming distinct subclones (single-cell examples in the right panel of **c**). ecDNAs amplifying *TERT* and *CDK4* are monotypic, likely reflecting temporal order among the events.

three variant ecDNA forms (e1-e3) exhibiting internal CN losses in distinct regions, consistent with derivative forms of e0 (**Fig. 4a, bottom left**). These variant ecDNAs coexisted with e0 in some cells, providing further support for their derivation from the parental molecule. Single-cell SV analysis identified specific rearrangements that account for the variant-specific internal deletions (**Fig. 4a, right**). We considered whether these deletions reflect active exclusion of deleterious genetic elements; in NCI-H69 and NCI-H524, the two samples where the variant ecDNAs are most prevalent, this seems unlikely as the deletions predominantly affect gene deserts or already-truncated genes (**Extended Data Fig. 12a-b**). Instead, the data are more consistent with ongoing ecDNA remodeling shaped by stochastic DNA damage and repair[30], with potential purifying selection preserving segments harboring essential oncogenic elements, such as *MYCN* (**Fig. 4b**). In addition, we identified cells exhibiting unique CN patterns in ecDNA regions, suggesting ongoing branched evolution (**Extended Data Fig. 12c**).

In contrast, GBM0510 provides evidence of convergent evolution[31] through ecDNA. In this tumor, `ECADeMix` revealed that three oncogenes—*TERT* (5p15.33), *EGFR* (7p11.2), and *CDK4*

370 (12q14.1)—were each amplified on separate ecDNA species (**Fig. 4c, top left** and **Extended Data**
371 **Fig. 13a**). *TERT* and *CDK4* ecDNAs were subclonal, present in 47% and 46% of cells,
372 respectively, and were structurally uniform within positive cells: *TERT* ecDNA arose from a local
373 chromothripsis event in the subtelomeric region 5p region, and *CDK4* ecDNA formed by a simple
374 head-to-tail rearrangement (**Fig. 4c, bottom left**). In contrast, *EGFR* amplification was clonal but
375 partitioned into three mutually exclusive ecDNA forms (e1-e3) across cells. The largest and most
376 prevalent (e1) was formed by a simple head-to-tail rearrangement, whereas e2 and e3 were
377 generated by distinct, complex rearrangements (**Fig. 4c, right**). The amplicon intervals for e1-e3
378 overlap only at the *EGFR* locus, and no SVs are shared among them, supporting independent
379 origins. Single-nucleotide variant (SNV) profiles (**Extended Data Fig. 13**) indicate that the
380 subclones defined by the three *EGFR* ecDNA forms share most SNVs and harbor few private
381 SNVs, consistent with relatively late emergence of these ecDNA forms[32]. Taken together, these
382 features support convergent evolution through recurrent, independent acquisition of *EGFR*-
383 amplifying ecDNAs. Compared with the multiple *EGFR* ecDNA species, *TERT* and *CDK4*
384 ecDNAs were structurally uniform, consistent with earlier acquisition, whereas their subclonality
385 may reflect stochastic loss from unequal ecDNA segregation (**Fig. 4d**). Similarly, in GBM0721
386 we identified three mutually exclusive ecDNA species amplifying *EGFR*, one of which
387 predominated (**Extended Data Fig. 13b**). In contrast, we found only a single ecDNA species for
388 *MDM4*, again supporting a late acquisition of ecDNA-mediated *EGFR* amplification in GBM.
389 Thus, these findings are consistent with ecDNAs in human cancers mediating both divergent and
390 convergent evolution, resulting in dynamic cell-to-cell diversification over time.

391

392 **ecDNA status by scWGS relative to bulk genome-graph predictions**

393 We next asked how single-cell analyses of ecDNA species and their evolutionary trajectories
394 compare to bulk genome-graph based predictions. For this comparison, we applied Amplicon
395 Architect (AA), a widely used method for identifying ecDNA from bulk sequencing data[33], to
396 pseudobulk data generated from our single-cell WGS (**Fig. 5a; Methods**). We then matched
397 HLAMP regions to AA-defined amplicons and compared their classifications. Importantly,
398 pseudobulk data derived from our scWGS are effectively analogous to bulk WGS but without
399 information on CN distributions over cells, enabling a head-to-head comparison of the two
400 approaches. At the case level, 9 out of 11 ecDNA-positive cases by single-cell CN distribution
401 were also called ecDNA-positive by AA (81% concordance). However, among 82 cases without a
402 CN distribution consistent with ecDNA, AA called 30 (37%) as ecDNA-positive. At the HLAMP
403 region level, stratified by cancer type (**Fig. 5b**), discrepancies were mainly in HGSOC and
404 hormone receptor-negative (HR−) breast cancers: these regions exhibited one or more narrow,
405 high CN peaks in single cells, indicating ICamp (median ecDNA score=0.129,

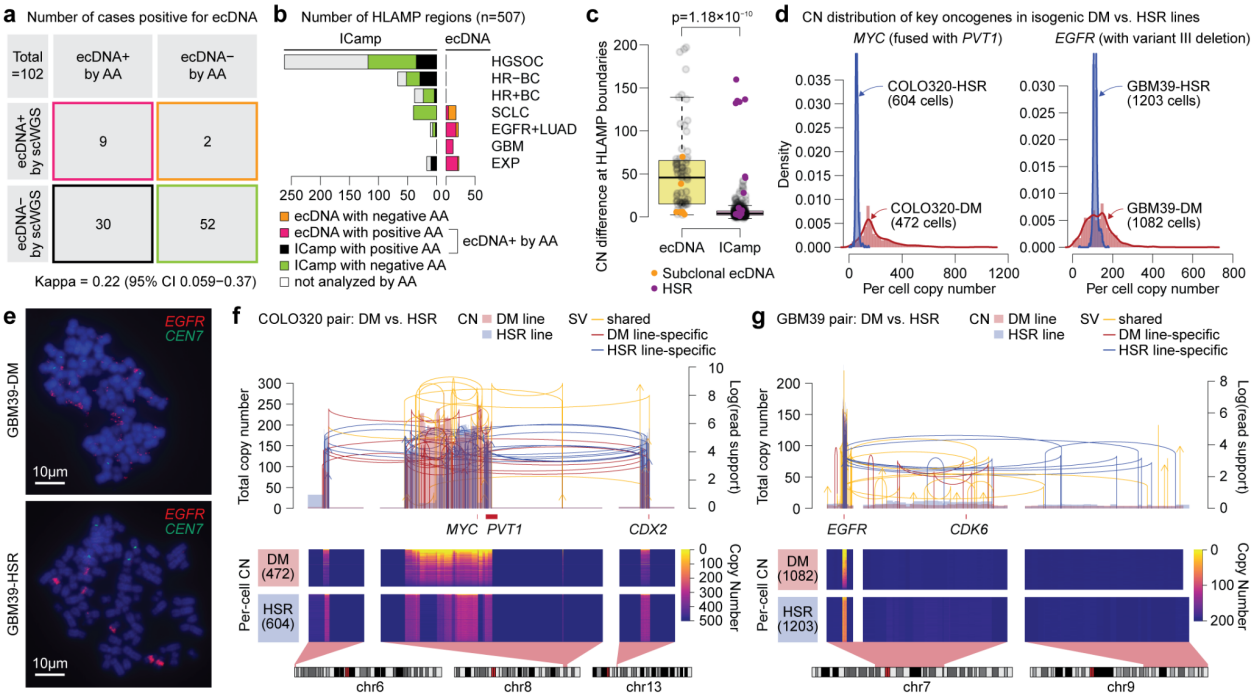


Fig. 5. scWGS distinguishes current ecDNA from historical bulk-level footprints. a, Case-level comparison of ecDNA predictions from bulk genome-graph methods (Amplicon Architect/Classifier) versus scWGS. **b**, HLAMP region-level comparison divided by cancer type. **c**, Junctional CN difference profiles in ecDNA and ICamp regions. Statistical significance was determined by Student's *t* test. **d**, CN distributions in isogenic paired cell lines: ecDNA-mediated double minutes (COLO320-DM and GBM39-DM) versus homogeneously staining regions (COLO320-HSR and GBM39-HSR). **e**, Cytogenetic features of GBM39 isogenic pair exhibiting numerous variable sized, dispersed signals consistent with *EGFR*-bearing ecDNA (top) and intrachromosomal clustered amplification manifesting as homogeneously staining region (HSR). **f-g**, Pseudobulk CN/SV profiles (top) and single-cell CN profiles (bottom) for DM versus HSR lines reveal genomic rearrangements that occurred after the most recent common ancestor and prior to HSR formation in COLO320 and GBM39 isogenic pairs.

range=0.040−0.380). In contrast, ecDNA-positive cases by AA in *EGFR*-mutated NSCLC, SCLC, and GBM mostly showed CN distributions consistent with ecDNAs, suggesting tissue-type specificity of ecDNA prevalence. Review of AA genome graphs suggested several potential sources of discrepancies (**Extended Data Fig. 14a-c**). In ecDNA-positive cases by scWGS, the CN transitions between unamplified and amplified segments were typically abrupt, consistent with circularization followed by amplification, and exhibited high junctional CN in pseudobulk profiles (**Fig. 5c** and **Extended Data Fig. 14a**)[34], with notable exceptions in cases with subclonal ecDNAs. However, in many HLAMP cases in HGSOC and HR− breast cancers, pseudobulk CN profiles showed stepwise increases in CN from unamplified to amplified regions across extended intervals, despite SVs arranged in cyclic paths (**Extended Data Fig. 14b**). This pattern is more consistent with progressive ICamp via multiple rearrangements, than with ecDNA[35].

429 Notably, three HLAMP cases showed circular paths in AA genome graphs and abrupt CN
430 transitions in pseudobulk profiles, a combination typical for ecDNA, yet their single-cell CN
431 distributions were instead consistent with ICamp (**Fig. 5c**). Two of these were isogenic cell line
432 models exhibiting homogeneously staining regions (HSRs) in DNA FISH (COLO320-HSR and
433 GBM39-HSR; **Fig. 5d-e**)[36,37]. Similar to their ecDNA-positive counterparts (COLO320-DM and
434 GBM39-DM, respectively), these HSR cell lines retain the highly amplified SV breakpoint
435 footprint of the initiating ecDNA (**Fig. 5f** and **5g**), indicating that the HSR was formed by
436 chromosomal integration of ecDNAs[38]. The third case was NCI-H82, a SCLC cell line, which was
437 labeled as ecDNA-positive in a previous study using AA[39]. The bulk genome graph showed a
438 cyclic path amplifying *MYC* with abrupt CN transitions, but the single-cell CN distribution was
439 sharply peaked, consistent with ICamp (**Extended Data Fig. 14c**). DNA FISH confirmed *MYC*
440 HSR formation in this line without evidence of ecDNA (**Extended Data Fig. 4b**).

441 Last, we analyzed genomic rearrangements associated with chromosomal integration of ecDNA
442 in isogenic pairs. In COLO320-HSR, the core amplicon contained several internal deletions
443 indicated by large, abrupt CN transitions, consistent with branched evolution from the parental
444 ecDNA prior to integration (**Fig. 5f**). In GBM39-HSR, HSR line-specific SVs included a
445 chromothripsis event linking the *EGFR* amplicon to the subtelomeric 9q region (**Fig. 5g**). These
446 HSR line-specific SVs inform mutational processes between the shared ecDNA ancestor and HSR
447 formation but do not, by themselves, identify the integrated chromosomes. Orthogonal mapping
448 from previous studies places the HSR on chrX for COLO320-HSR[40]. Together, these analyses
449 show that single-cell CN distributions can distinguish actively maintained ecDNAs from historical
450 ecDNA footprints re-integrated into chromosomes, substantially revise bulk genome-graph
451 predictions, and place HLAMP events within their evolutionary context.

452

453 **Discussion**

454 Leveraging scWGS, we quantified cell-to-cell variability in HLAMP across four major cancer
455 types and experimental models, and linked this variability to specific generative and modulatory
456 mechanisms. In contrast to bulk genome-graph predictions, single-cell CN distributions of
457 HLAMP events distinguish episomal ecDNA from ICamp, including HSRs derived from
458 chromosomal re-integration of ecDNAs. Although cyclic genome graphs in bulk WGS can
459 nominate candidate ecDNAs, many candidates, particularly in HGSOC and HR− breast cancers,
460 show narrow, sharply peaked CN distributions consistent with ICamp. In these tumor types,
461 HLAMP heterogeneity appears largely shaped by ICamp evolution through numeric modulation
462 and structural remodeling, which can generate substantial intratumoral heterogeneity in a
463 punctuated manner. Notably, subclones that lose oncogene amplification through aneuploidy

464 warrant further clinical investigation, as selection for oncogene-low subclones could promote early
465 resistance to oncogene-directed targeted therapies[41].

466 In contrast to enrichment-based single-cell methods that selectively profile ecDNA[42–44], scWGS
467 enables genome-wide characterization of HLAMP evolution in single cancer cells, including both
468 ecDNAs and ICamps. Compared with other genome-wide single-cell profiling methods such as
469 single-cell ATAC[45] and RNA sequencing[46], which are being explored for ecDNA detection,
470 DLP+-based scWGS or related technologies[47,48] provide unbiased CN profiles and SV information
471 with high CN fidelity and nucleotide-level coverage at HLAMP loci in single cells, both of which
472 are crucial for defining the mechanisms and evolution of HLAMP events. Heuristically, in the
473 analyses presented here, we required at least 30 high-quality single-cell CN profiles for ecDNA
474 detection. However, we suggest further algorithmic improvements and context-specific events
475 with ultra-high copies could be resolved to fewer than 30 cells.

476 Our results suggest that ecDNA is prevalent in select cancer types, in contrast to recent studies
477 based on bulk genome-graph predictions[49,50]. In cancer types where ecDNA has been extensively
478 described, namely GBM and SCLC, our scWGS-based analysis confirms a high prevalence of
479 ecDNA. We also identified 3 ecDNA species in 2 of 5 *EGFR*-mutated NSCLCs, amplifying key
480 oncogenes including *EGFR*, *MDM2*, and *NKX2-1*. The contribution of ecDNAs to the biology of
481 *EGFR*-mutated lung cancer and to responses to targeted therapies will be of clinical interest. By
482 contrast, we did not observe any HLAMP event with a CN distribution consistent with ecDNA
483 among 72 patients with HGSOC or HR− breast cancer. Although cyclic paths were identified
484 around HLAMP regions in many cases, their single-cell CN distributions and SV features were
485 instead consistent with ICamps arising through multi-step rearrangement processes. In some cases,
486 rare cells displayed CN outliers superimposed on a global CN distribution consistent with ICamp.
487 These outlier cells may contain ecDNA, but their frequency was exceedingly low. The striking
488 paucity of ecDNA in HGSOC and HR− breast cancers suggests that these tissue contexts could be
489 less permissive to ecDNA formation or maintenance. In addition, previous studies indicated that
490 anticancer therapies, such as chemotherapy or radiotherapy, can induce ecDNA formation[51,52],
491 potentially reshaping ecDNA prevalence in patients with advanced cancer after treatment. In line
492 with this, a clinical study in patients with platinum-resistant HGSOC reported double minutes in
493 cancer cells from patients' ascites[53]. The interplay between tissue context and therapy-induced
494 genome instability in shaping the HLAMP landscape warrants further investigation.

495 Motivated by longstanding efforts to develop ecDNA-directed therapeutics[13–15,54,55] and growing
496 recognition of the profound impact of ecDNA on cancer evolution, there is increasing interest in
497 detecting ecDNAs in patient biospecimens. Conventional cytogenetic techniques like FISH require
498 metaphase cells to accurately detect ecDNAs[56], but these are not often available for clinical tumor
499 samples. Instead, high-throughput sequencing of clinical tumors has become an integral part of

500 oncology care, and the use of such assays to detect ecDNA has been proposed[8]. Our findings
501 highlight intrinsic challenges to this approach. A key genomic hallmark of prevalent ecDNA is
502 pronounced cell-to-cell CN variability, which cannot be resolved from single-sample bulk
503 sequencing. Genome graph-based approaches instead depend on sensitive detection of SVs from
504 WGS and cannot reliably distinguish extant ecDNA from its vestigial footprints after chromosomal
505 integration. Given the dynamic nature of HLAMP evolution, including the widespread punctuated
506 modulation of ICamp events observed in our study, scalable scWGS, or targeted single cell
507 approaches whereby cellular distributions can be read out on specific loci, may offer a more
508 promising path for clinical ecDNA detection and monitoring in the future.

509 In conclusion, our analysis of >86,000 single cell cancer genomes defines interpretable CN
510 distributions of HLAMP events consistent with distinct ICamp and ecDNA generative processes
511 and elucidates how these mechanisms diversify cancer cell populations. These dynamics extend
512 beyond ecDNA and are pervasive in tumors with ICamp, underscoring ongoing genome
513 remodeling. Single-cell genomic profiling of longitudinal clinical specimens, integrated with
514 orthogonal readouts of gene expression[48], would enable direct measurement of how therapy
515 reshapes HLAMP evolution. The derivative single-cell readouts of architecture, abundance and
516 clonal selection in cancer cells will therefore help to identify actionable dependencies and inform
517 therapeutic strategies through an evolutionary lens.

Acknowledgments

This work was supported by the Halvorsen Center for Computational Oncology and Cycle for Survival supporting Memorial Sloan Kettering Cancer Center. S.P.S. holds the Nicholls Biondi Chair in Computational Oncology and is a Susan G. Komen Scholar. Additional support from the Department of Defense Congressionally Directed Medical Research Program award [W81XWH-20-1-0565], an Ovarian Cancer Research Alliance (OCRA) Collaborative Research Development Grant [648007], an NIH R01 CA281928-01, and by the Seidenberg Family Foundation funds to SPS. Additional support was provided through the MSK Core Grant (P30 CA008748), the National Cancer Institute of the National Institute of Health (K08CA301011 to JJL; K99CA277562 to SS; R01CA282913 to AV), the Burroughs Wellcome Fund Career Award for Medical Scientists (to JJL), the Cancer Grand Challenge Initiative (to AV), the American Cancer Society Discovery Boost Grant (to AV), the Mark Foundation for Cancer Research ASPIRE Grant (to AV). We thank Gryte Satas, Ignacio Vazquez-Garcia, Samuel Freeman, Matthew Zatzman, Alessandro Grande, Giovanni Santacatterina, and the rest of the Shah, McPherson, and Lee labs for discussions and technical advice.

Author Contributions

JJL and SPS conceived the project and supervised study design. JJL and SS designed experiments and performed primary computational analyses, with conceptual input from MJW and AM. SS led development of the amplification classifier `eicicle`. MAM led development of the ecDNA deconvolution algorithm `ECADeMix`. SS performed single-cell RNA sequencing analyses. MJW conducted allele-specific copy-number analyses. MY, DHA, EGS, and AV performed cytogenetic experiments. MY and JL conducted cell culture experiments. PM, ER, AQV, and CMR generated data from patient-derived xenograft models. PR, SA, N. Rekhtman, VT, MMA, HAY, KKHY, and CMR generated data from clinical tumor samples. JJL, SS, MAM, MJW, KT, CT, CZJP, SM, and AM performed secondary computational analyses. SC, EH, and APS established and maintained computational analysis pipelines. MW and NM supervised single-cell whole-genome sequencing procedures. JJL, SS, MAM, N. Rusk, and SPS wrote the original draft, and all authors revised and approved the final manuscript.

Competing Interests

A patent application related to the determination of ecDNA versus ICamp based on single-cell DNA CN information has been filed by Memorial Sloan Kettering Cancer Center. AQV is a current employee of AstraZeneca. PR reports research funding from Grail, Novartis, AstraZeneca, EpicSciences, Invitae/ArcherDx, Biothernostics, Tempus, Neo- genomics, Biovica, Guardant, Personalis, Myriad and consulting or advisory role for Novartis, AstraZeneca, Pfizer, Lilly/Loxo, Prelude Therapeutics, Epic Sciences, Daiichi-Sankyo, Foundation Medicine, Inivata, Natera, Tempus, SAGA Diagnostics, Paige.ai, Guardant, and Myriad, outside the scope of this work. SA

554 is cofounder and shareholder of Genome Therapeutics and reports advisory roles for Chordia
555 Therapeutics and Sangamo Therapeutics, outside the scope of this work. MMA reports research
556 funding from Amgen, AstraZeneca, Bristol Myers Squibb, Genentech and Lilly, consulting or
557 advisory roles for Affini-T, AstraZeneca, Blueprint Medicines Corporation, D3Bio, EMD Serono,
558 Gritstone, Iovance, Merck, Merus, Mirati, Novartis, Pfizer and Synthekine, and serves on the Data
559 Safety Monitoring Boards for Apollomics and Bristol Myers Squibb, outside the scope of this
560 work. HAY reports consulting or advisory role for Takeda, Taiho, Black Diamond, BMS, AbbVie,
561 Amgen, AstraZeneca, Daiichi Sankyo, Ipsen, and Pfizer and serves on the Data and Safety
562 Monitoring Board for Janssen and Mythic Therapeutics, outside the scope of this work. CMR
563 reports consulting or advisory role for Genentech/Roche, AstraZeneca, Amgen, Jazz
564 Pharmaceuticals, Earli, AbbVie, Daiichi Sankyo/UCB Japan, Merck, Auron Therapeutics, DISCO
565 and research funding from Merck, Genentech/Roche, Daiichi Sankyo, outside the scope of this
566 work. SPS reports research funding from AstraZeneca and Bristol Myers Squibb, outside the scope
567 of this work.

568 **Data Availability**

569 Newly generated raw scWGS data will be available from the NCBI dbGaP prior to publication.
570 Previously published raw scWGS datasets are available through European Genome-Phenome
571 Archive (references 17 and 19: EGAS00001006343 and EGAS00001003190) and dbGaP under
572 the National Institute of Health (references 20 and 21: phs002857.v3.p1).

573 **Code Availability**

574 Source code for the computational models and scripts used for analysis in this manuscript can be
575 found at https://github.com/shahcompbio/hlamp.

576 **Figure Legends**

577 **Fig. 1. Cell-to-cell copy-number variance in scWGS differentiates ecDNA-mediated**
578 **amplifications from intrachromosomal amplifications. a,** Cancer types and case counts (top)
579 and schematic illustration of study design (bottom). Organ and model illustrations are from
580 BioRender. **b,** Number of high-quality cancer cells analyzed (x-axis, log-scale) by tumor type (y-
581 axis). **c,** Frequently amplified oncogenes and their case counts. **d,** Inheritance patterns for
582 intrachromosomal amplification (ICamp) versus extrachromosomal circular DNA (ecDNA; left)
583 and expected CN distributions across cells (right). **e,** Representative examples of ICamp and
584 ecDNA, from a patient with EGFR-mutated lung adenocarcinoma (left) and glioblastoma (right).
585 **f,** Conceptual illustration of two metrics used to classify HLAMP events. **g,** Classification of 507
586 HLAMP regions from 86,239 cancer cells based on single-cell CN distribution. **h,** Fraction of
587 clonal and subclonal HLAMP regions by mechanism, across cases.

588

589 **Fig. 2. Contrasting mechanisms of ICamp CN heterogeneity by numeric and structural**
590 **modulators. a,** Prevalence of multi-peak CN distribution among ICamp cases. **b,** An ICamp case
591 exhibiting multi-peak CN distribution across cells, indicating subclones with distinct *ERBB2* CN.
592 **c,** CN variance of HLAMP events stratified by initiating mechanism (BFB, breakage-fusion
593 bridge; CGR-NOS, complex genomic rearrangement, not otherwise specified; CGR-TRA,
594 complex genomic rearrangement, translocation-rich). **d,** Initiating and modulating mechanisms for
595 ICamps. **e,** Pairwise comparisons of HLAMP-associated arm-level CN profiles between
596 subclones. **f,** A classical example of aneuploidy-mediated doubling of *CCNE1* amplification in
597 EL011. Compared to clone C, all the HLAMP segments exhibit proportional CN increase in clone
598 H. **g,** A structural modulation of 17q amplification in OV-002. Clone B exhibits further
599 amplification of two segments (one amplifying *MIR21*) within the HLAMP segment in clone A.
600 **h,** Multiple CN peaks of *CCND1* amplification in OV-052 matching the subclonal phylogeny
601 (left), clone-specific CNs and SVs indicative of BFB-mediated amplification and subsequent
602 modulation by arm-level aneuploidy (middle), and schematics of the mechanism (right). **i,** UMAP
603 embedding of 2,244 cancer cells from scRNA sequencing of OV-052 (left) and summary CN
604 profiles of inferred clones in the malignant compartment (right). A subclone with no *CCND1*
605 amplification (cluster 3; n=431 cells) is detected. **j,** Normalized expression level of *CCND1* in
606 each cluster. **k,** Gene set enrichment analysis comparing cluster 3 with the other clusters. The top
607 8 enriched pathways are shown.

608

609 **Fig. 3. Resolving mechanisms of segregation of co-amplified oncogenes in ecDNA in single**
610 **cells. a-b,** Genome-wide CN profiles (left) and CN correlations between key oncogenes (right) for

611 GBM0721 and EL001. **c-d,** Single cell CN/SV profiles indicating two distinct simple ecDNAs in
612 GBM0721 and two distinct hybrid ecDNAs in EL001. **e,** Hierarchical clustering of HLAMP
613 regions in EL001 based on pairwise CN correlations reveals two ecDNA species co-amplifying
614 oncogenes. **f,** Decomposition of ecDNA heteroplasmy in EL001 by ECADeMix. HLAMP
615 segments used (S1-S4) are indicated in d. **g,** Pseudobulk CN/SV profile from Lx516 showing
616 multi-chromosomal chromothripsis leading to HLAMP of *KDM2A*. **h,** Clustering based on
617 pairwise CN correlation identifies a hybrid ecDNA and a subclonal ICamp event. Rearrangements
618 between segments are illustrated as arcs on the top. Chromosomal origins in the left heatmap.
619 Subclonal CN and ecDNA scores on the right side. Inset plots show the reconstructed ecDNA
620 structure (left) and a representative CN correlation between two segments that comprise the
621 ecDNA. **i,** Schematic illustration of HLAMP mechanisms in Lx516.

622

623 **Fig. 4. Modes of ongoing ecDNA evolution. a,** Chromothripsis-derived ecDNA amplifying
624 *MYCN* in the NCI-H69 SCLC cell line (top left); single-cell CN profiles of the amplicon and
625 ecDNA heteroplasmy delineated by ECADeMix (bottom left). CN/SV profiles from 4 cancer cells
626 with distinct ecDNA subspecies (right). e0 is the likely parental ecDNA; e1-e3 are daughter
627 ecDNAs derived from e0 via internal rearrangements. **b,** Schematic illustration of branched
628 evolution via ecDNA rearrangement. **c-d,** Pseudobulk genome-wide CN profile of glioblastoma
629 GBM0510, exhibiting multiple ecDNA-mediated HLAMPs (**c,** top left); single-cell CN profiles of
630 the amplicons and decomposed ecDNA species (**c,** bottom left). *EGFR* is amplified by three
631 distinct ecDNAs (e1, e2, and e3) in a mutually exclusive manner across cells, forming distinct
632 subclones (single-cell examples in the right panel of **c**). ecDNAs amplifying *TERT* and *CDK4* are
633 monotypic, likely reflecting temporal order among the events.

634

635 **Fig. 5. scWGS distinguishes current ecDNA from historical bulk-level footprints. a,** Case-
636 level comparison of ecDNA predictions from bulk genome-graph methods (Amplicon
637 Architect/Classifier) versus scWGS. **b,** HLAMP region-level comparison divided by cancer type.
638 **c,** Junctional CN difference profiles in ecDNA and ICamp regions. Statistical significance was
639 determined by Student's *t* test. **d,** CN distributions in isogenic paired cell lines: ecDNA-mediated
640 double minutes (COLO320-DM and GBM39-DM) versus homogeneously staining regions
641 (COLO320-HSR and GBM39-HSR). **e,** Cytogenetic features of GBM39 isogenic pair exhibiting
642 numerous variable sized, dispersed signals consistent with *EGFR*-bearing ecDNA (top) and
643 intrachromosomal clustered amplification manifesting as homogeneously staining region (HSR).
644 **f-g,** Pseudobulk CN/SV profiles (top) and single-cell CN profiles (bottom) for DM versus HSR

645 lines reveal genomic rearrangements that occurred after the most recent common ancestor and
646 prior to HSR formation in COLO320 and GBM39 isogenic pairs.

647 **References**

648 1. Slamon, D. J. *et al.* Use of Chemotherapy plus a Monoclonal Antibody against HER2 for
649 Metastatic Breast Cancer That Overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792 (2001).

650 2. Bang, Y.-J. *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy
651 alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction
652 cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* **376**, 687–697
653 (2010).

654 3. Yagisawa, M. *et al.* Trastuzumab deruxtecan in advanced solid tumors with human
655 epidermal growth factor receptor 2 amplification identified by plasma cell-free DNA
656 testing: A multicenter, single-arm, phase II basket trial. *J. Clin. Oncol.* **42**, 3817–3825
657 (2024).

658 4. Regitnig, P., Schippinger, W., Lindbauer, M., Samonigg, H. & Lax, S. F. Change of HER-
659 2/neu status in a subset of distant metastases from breast carcinomas. *J. Pathol.* **203**, 918–
660 926 (2004).

661 5. Seol, H. *et al.* Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its
662 clinicopathological significance. *Mod. Pathol.* **25**, 938–948 (2012).

663 6. Jordan, N. V. *et al.* HER2 expression identifies dynamic functional states within circulating
664 breast cancer cells. *Nature* **537**, 102–106 (2016).

665 7. Filho, O. M. *et al.* Impact of HER2 heterogeneity on treatment response of early-stage
666 HER2-positive breast cancer: Phase II neoadjuvant clinical trial of T-DM1 combined with
667 pertuzumab. *Cancer Discov.* **11**, 2474–2487 (2021).

668 8. Weiser, N. E., Watkins, T. B. K., Chang, H. Y. & Mischel, P. S. A guide to
669 extrachromosomal DNA: Cancer's dynamic circular genome. *Cancer Discov.* **15**, 1105–
670 1114 (2025).

671 9. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and
672 genetic heterogeneity. *Nature* **543**, 122–125 (2017).

673 10. Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression.
674 *Nature* **600**, 731–736 (2021).

675 11. Snuderl, M. *et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in
676 glioblastoma. *Cancer Cell* **20**, 810–817 (2011).

677 12. Lange, J. T. *et al.* The evolutionary dynamics of extrachromosomal DNA in human cancers.
678 *Nat. Genet.* **54**, 1527–1533 (2022).

679 13. Clarke, T. L. *et al.* Histone lysine methylation dynamics control *EGFR* DNA copy-number
680 amplification. *Cancer Discov.* **10**, 306–325 (2020).

681 14. Tang, J. *et al.* Enhancing transcription-replication conflict targets ecDNA-positive cancers.
682 *Nature* **635**, 210–218 (2024).

683 15. Dharanipragada, P. *et al.* Blocking Genomic Instability Prevents Acquired Resistance to
684 MAPK Inhibitor Therapy in Melanoma. *Cancer Discov.* **13**, 880–909 (2023).

685 16. Laks, E. *et al.* Clonal decomposition and DNA replication states defined by scaled single-
686 cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).

687 17. Funnell, T. *et al.* Single-cell genomic variation induced by mutational processes in cancer.
688 *Nature* **612**, 106–115 (2022).

689 18. Haigis, K. M., Cichowski, K. & Elledge, S. J. Tissue-specificity in cancer: The rule, not the
690 exception. *Science* **363**, 1150–1151 (2019).

691 19. Salehi, S. *et al.* Clonal fitness inferred from time-series modelling of single-cell cancer
692 genomes. *Nature* **595**, 585–590 (2021).

693 20. McPherson, A. *et al.* Ongoing genome doubling shapes evolvability and immunity in
694 ovarian cancer. *Nature* **644**, 1078–1087 (2025).

695 21. Williams, M. J. *et al.* Tracking clonal evolution during treatment in ovarian cancer using
696 cell-free DNA. *Nature* **647**, 757–765 (2025).

697 22. Pradella, D. *et al.* Engineered extrachromosomal oncogene amplifications promote
698 tumorigenesis. *Nature* **637**, 955–964 (2025).

699 23. Lee, J. J.-K. *et al.* ERα-associated translocations underlie oncogene amplifications in breast
700 cancer. *Nature* **618**, 1024–1032 (2023).

701 24. Doob, J. L. Topics in the theory of Markoff chains. *Trans. Amer. Math. Soc.* **52**, 37–64
702 (1942).

703 25. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution
704 of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).

705 26. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**,
706 112–121 (2020).

707 27. Luo, Y. *et al.* Neoadjuvant PARPi or chemotherapy in ovarian cancer informs targeting
708 effector Treg cells for homologous-recombination-deficient tumors. *Cell* **187**, 4905–
709 4925.e24 (2024).

710 28. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers.
711 *Nature* **396**, 643–649 (1998).

712 29. Kim, M. *et al.* Single-cell mtDNA dynamics in tumors is driven by coregulation of nuclear
713 and mitochondrial genomes. *Nat. Genet.* **56**, 889–899 (2024).

714 30. Ilić, M., Zaalberg, I. C., Raaijmakers, J. A. & Medema, R. H. Life of double minutes:
715 generation, maintenance, and elimination. *Chromosoma* **131**, 107–125 (2022).

716 31. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by
717 multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

718 32. Ozawa, T. *et al.* Most human non-GCIMP glioblastoma subtypes evolve from a common
719 proneural-like precursor glioma. *Cancer Cell* **26**, 288–300 (2014).

720 33. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using
721 AmpliconArchitect. *Nat. Commun.* **10**, 392 (2019).

722 34. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event
723 during cancer development. *Cell* **144**, 27–40 (2011).

724 35. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic
725 rearrangement patterns. *Nature* **538**, 378–382 (2016).

726 36. Alitalo, K., Schwab, M., Lin, C. C., Varmus, H. E. & Bishop, J. M. Homogeneously
727 staining chromosomal regions contain amplified copies of an abundantly expressed cellular
728 oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc.*
729 *Natl. Acad. Sci. U. S. A.* **80**, 1707–1711 (1983).

730 37. Nathanson, D. A. *et al.* Targeted therapy resistance mediated by dynamic regulation of
731 extrachromosomal mutant EGFR DNA. *Science* **343**, 72–76 (2014).

732 38. Rosswog, C. *et al.* Chromothripsis followed by circular recombination drives oncogene
733 amplification in human cancer. *Nat. Genet.* 1–13 (2021).

734 39. Zhang, J. *et al.* Amplification of extrachromosomal *MYC* paralogs shapes
735 immunosuppressive tumor microenvironment in small cell lung cancer. *Clin. Cancer Res.*
736 **31**, 4529–4542 (2025).

737 40. Nichols, A. *et al.* Chromosomal tethering and mitotic transcription promote ecDNA nuclear
738 inheritance. *Mol. Cell* **85**, 2839–2853.e8 (2025).

739 41. Chen, W. *et al.* Trastuzumab Deruxtecan resistance via loss of HER2 expression and
740 binding. *Cancer Discov.* **16**, 235–249 (2026).

741 42. Hung, K. L. *et al.* Targeted profiling of human extrachromosomal DNA by CRISPR-
742 CATCH. *Nat. Genet.* **54**, 1746–1754 (2022).

743 43. Chapman, O. S. *et al.* Circular extrachromosomal DNA promotes tumor heterogeneity in
744 high-risk medulloblastoma. *Nat. Genet.* **55**, 2189–2199 (2023).

745 44. Chen, J. P. *et al.* scCircle-seq unveils the diversity and complexity of extrachromosomal
746 circular DNAs in single cells. *Nat. Commun.* **15**, 1768 (2024).

747 45. Jiang, R. *et al.* scCirclehunter delineates ecDNA-containing cells using single-cell ATAC-
748 seq, with a focus on glioblastoma. *Cell Discov.* **11**, 98 (2025).

749 46. Hendriksen, J. D. *et al.* Resolving clonal evolution and selection of extrachromosomal DNA
750 at single-cell resolution. *Genome Biol.* **27**, 10 (2026).

751 47. Wang, K. *et al.* Archival single-cell genomics reveals persistent subclones during DCIS
752 progression. *Cell* **186**, 3968–3982.e15 (2023).

753 48. Wang, K. *et al.* Coalescing single-cell genomes and transcriptomes to decode breast cancer
754 progression. *Cell* **188**, 6355–6369.e16 (2025).

755 49. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor
756 outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).

757 50. Bailey, C. *et al.* Origins and impact of extrachromosomal DNA. *Nature* **635**, 193–200
758 (2024).

759 51. Coquelle, A., Toledo, F., Stern, S., Bieth, A. & Debatisse, M. A new role for hypoxia in
760 tumor progression. *Mol. Cell* **2**, 259–265 (1998).

761 52. Shoshani, O. *et al.* Chromothripsis drives the evolution of gene amplification in cancer.
762 *Nature* **591**, 137–141 (2021).

763 53. Raymond, E. *et al.* Effects of hydroxyurea on extrachromosomal DNA in patients with
764 advanced ovarian carcinomas. *Clin. Cancer Res.* **7**, 1171–1180 (2001).

765 54. Von Hoff, D. D. *et al.* Elimination of extrachromosomally amplified MYC genes from
766 human tumor cells reduces their tumorigenicity. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 8165–
767 8169 (1992).

768 55. Shimizu, N., Itoh, N., Utiyama, H. & Wahl, G. M. Selective entrapment of
769 extrachromosomally amplified DNA by nuclear budding and micronucleation during S
770 phase. *J. Cell Biol.* **140**, 1307–1320 (1998).

771 56. Hung, K. L., Mischel, P. S. & Chang, H. Y. Gene regulation on extrachromosomal DNA.
772 *Nat. Struct. Mol. Biol.* **29**, 736–744 (2022).

773

Methods

Study Cohort

Our study cohort comprises both previously published single-cell whole-genome sequencing (scWGS) datasets from our group and collaborators and newly generated scWGS data produced for this study. All scWGS data were generated using the direct library preparation plus (DLP+) platform.

For published data, we assembled a meta-analysis cohort from 4 previous studies[17, 19-21], including surgically resected high-grade serous ovarian cancer (HGSOC) samples from 46 patients, patient-derived xenograft (PDX) models of HGSOC from 16 patients, and PDX models of hormone receptor-negative breast cancers from 10 patients (9 were also negative for *HER2* amplification and 1 *HER2*-amplified). PDX-based scWGS data was generated at the British Columbia Cancer Research Centre, whereas scWGS data from clinical HGSOC tumor samples were generated at Memorial Sloan Kettering Cancer Center (MSKCC).

We additionally generated new DLP+-based scWGS data for this study from 3 cancer subtypes—small-cell lung cancer (SCLC), *EGFR*-mutated non-small cell lung cancer (NSCLC), and glioblastoma (GBM)—as well as several experimental models. For SCLC, we profiled PDX models established from 5 patients and rapid autopsy-based tumors from 3 participans. For *EGFR*-mutated NSCLC, we used frozen surgical tumor resections from 5 patients. For GBM, we used freshly dissociated single-cell suspensions derived from surgical resection from 4 patients. These samples were processed at the Integrated Genomics Operation at Memorial Sloan Kettering Cancer Center.

All patients provided informed consent under the Institutional Review Board (IRB)-approved clinical study protocols. The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice (GCP) guidelines.

DLP+ Analyses

<u>ScWGS Data Alignment and Quality Control</u>

We processed scWGS data as described previously using the publicly available `Mondrian` software suite (`https://github.com/mondrian-scwgs/mondrian`), executed within the Isabl framework[57]. After trimming adaptor sequences, reads from each cell were aligned to the human reference genome GRCh37 using bwa-mem v0.7.17. PCR duplicates were identified and marked using Picard v2.27.4. Alignment and quality metrics were collected from each BAM file using `Picard` (InsertWgsMetrics and WgsMetrics) and `samtools` flagstat. BAM files from each cell were then merged to generate a pseudobulk BAM file.

For first-pass inference of single-cell copy-number profiles, reads were counted in non-overlapping 500-Kbp bins genome-wide for each cell. GC-content bias was adjusted using modal regression normalization. Copy-number states were then inferred using `HMMcopy`[58] by selecting the best-fit model.

Finally, we integrated alignment metrics and copy-number profiles through the quality-control pipeline to estimate quality of each cell and to flag S-phase cells and doublets. Based on this result, we filtered non-cancer cells (either control cells or those exhibiting euploid copy-number profiles), low-quality cells (quality score less than 0.75), cells with low read count (total mapped reads ≤ 250,000), S-phase replicating cells, and doublets, resulting in 86,239 high-quality single-cell cancer genomes.

Somatic Structural Variant Analysis

We detected somatic structural variants (SVs) using `GRIDSS` v2.13.2[59] with default parameters and hg19 blacklisted regions from the ENCODE consortium (ENCFF001TD0; https://www.encodeproject.org/files/ENCFF001TDO/). For the 82 cases where bulk whole-genome sequencing (WGS) of normal tissue is available, we ran GRIDSS in multi-sample mode, treating all DLP+ libraries as tumor samples and the matched normal bulk WGS as the normal. For the remaining 18 cases without matched normal WGS (**Table S1**), we ran GRIDSS in tumor-only mode. VCF output files were post-processed to collapse multi-sample SV calls into a single composite column. The resulting composite VCF was then subject to GRIPSS v.2.3 (https://github.com/hartwigmedical/hmftools/tree/purple-v3.7.2/gripss) for somatic filtering. GRIPSS was run with default settings, using the panel of normals, fusion hotspot, and repeat information recommended by the Hartwig Medical Foundation tool suite. The resultant VCF file was converted into BEDPE format for further downstream analysis (associated scripts are available at: https://github.com/shahcompbio/hlamp/tree/main/src/structural_variant_calling).

We next genotyped the SVs detected at pseudobulk level in each single cell using a modified version of `SVtyper` (https://github.com/marcjwilliams1/svtyper)[60], as described in our recent study[21]. In brief, we modified `SVtyper` to compute the number of read support in each single cell. It provides per-cell read support for the reference and variant sequences, including split reads that directly capture the breakpoint and discordant read pairs with larger-than-expected insert sizes with proper orientation (+/−) or those in variant orientation (−/+, −/−, +/+) or mapping to different chromosomes. We added features to output cell IDs as concatenated string in the VCF and generate matrices linking cell IDs and SV IDs for downstream analysis.

SV Breakpoint-Concordant Single-Cell Copy-Number Analysis

To infer the generative mechanisms of high-level copy-number amplification (HLAMP) events from associated SV features, we generated case-specific genomic segments by integrating SV breakpoints with a genome-wide tiling of non-overlapping 2 Mbp bins. Using pseudobulk BAM file, we then computed per-cell read counts for each segment in all high-quality cancer cells (https://github.com/shahcompbio/hlamp/tree/main/src/sv_aware_seg_cn/02_sv-seg-cn/). Then, we computed the average reads per copy per base in the HMMCopy 500kb copy number profiles and used this constant for each cell to scale the segment read counts to

approximate segment copy numbers (https://github.com/shahcompbio/hlamp/tree/main/src/sv_aware_seg_cn/03_calibrate_cn_gc_cell_inclusion.py). This scaling ensures that despite varying segment sizes, the approximate copy numbers would be comparable between segments. Next, we applied GC correction via modal quantile regression[16]. Finally, we scaled these GC-corrected values to match the cell ploidy estimated using HMMCopy (assuming that any HLAMPs whose copy number is now captured more accurately would not meaningfully contribute to the cell ploidy due to their small size).

Clustering Cancer Cells to Identify Subclones

High-quality cancer cells were clustered based on their genome-wide copy number profiles. Briefly, we first restricted the bin set to informative loci by removing bins with zero normalized read coverage, excluding short segments (<1000 bp), and removing bins overlapping centromeric regions. We then constructed a cell by bin matrix from the GC-corrected copy number layer, set any non-finite values to zero, and standardized features per bin using z-scoring with clipping of extreme values (maximum absolute value 10). Dimensionality reduction was performed by principal component analysis (PCA; up to 50 components), followed by construction of a k-nearest neighbor graph (k=15) in PCA space and community detection using the Leiden algorithm at resolution 0.7 (unless otherwise specified). These subclones were labeled by the Leiden partition. This procedure was followed by generating case-specific diagrams (https://github.com/shahcompbio/hlamp/tree/main/src/sv_aware_seg_cn/04_dlp.composite.plot.v4.R) including pseudobulk copy number plot, per-cell copy number heatmap, and subclone-specific copy-number plots for manual review.

Identifying HLAMP regions

Using signals package[17], we first estimated pseudobulk and subclonal copy-numbers based on the SV breakpoint-concordant single-cell copy-number profiles and subclones defined above. Then we established baseline copy number for each chromosome arm at pseudobulk level, by selecting the copy-number state with the greatest combined segmental length, with the lowest possible baseline copy number value being 2, excluding acrocentric arms (13p, 14p, 15p, 21p, and 22p). We then marked the segments of which copy number value is more than 3 times of the arm-level copy-number baseline in at least one subclone and combined contiguous segments to define amplified regions. For those regions that were bordered by fold-back inversions (SVs of which orientation is −/− or +/+ and the distance between two breakpoints were 5000 bp or less), we calibrated the borders by including both inversion breakpoints (https://github.com/shahcompbio/hlamp/tree/main/src/sv_aware_seg_cn/05_dlp.oncogene.cn.v4.R and https://github.com/shahcompbio/hlamp/tree/main /src/sv_aware_seg_cn/06_dlp.region.cn.v4.R). Applying this process on 102 cases identified 2,925 amplified regions. Among these, 507 segments of which weighted mean copy number is 6 or greater, we labeled them as HLAMP and used as main data for downstream analyses.

117

118 Mechanistic classification of HLAMP regions based on SV features

119 SVs at boundaries between amplified and un-amplified regions can provide insights into the
120 generative mechanisms of an amplicon[23,61]. Our SV breakpoint-aware single-cell copy-number
121 profiles enabled classification of HLAMP events into mechanistic classes based on SV patterns at
122 HLAMP boundaries.

123 We first clustered the SV breakpoints using clusterSV (https://github.com/ cancerit/
124 ClusterSV), which accounts for breakpoint proximity, background SV rates, and SV type-
125 specific size distributions, as described previously[26]. We defined chromothripsis as an SV cluster
126 meeting three criteria: i) ≥20 SV breakpoints on a single chromosome; ii) SVs represented in all 4
127 read orientations, and iii) unbiased representation across the 4 orientation categories, assessed by
128 permutation testing. For iii), we tested whether the observed counts of SV breakpoints across 4
129 orientations differed from expectation under equal probability using a chi-square goodness-of-fit
130 test (implemented via the `permutation_test` function in https://github.com/
131 shahcompbio/hlamp/tree/main/src/sv_aware_seg_cn/07_ amp.mechanism.R).

132 HLAMP events whose boundary SVs were part of chromothripsis cluster were classified as
133 **chromothripsis**. HLAMP events bordered by fold-back inversions (FBIs) were classified as **FBI**.
134 If both boundaries were connected by a single head-to-tail SV, the event was classified as **tandem**
135 **duplication**. If neither boundary had a detected SV, or if boundaries were supported only by single
136 breakends, the event was classified as **unknown**.

137 Remaining HLAMP regions were typically embedded within complex genomic rearrangement
138 (CGR) clusters that did not meet the chromothripsis criteria. Within these, we additionally assigned
139 an FBI class to clusters with ≥10 SVs in which FBI were the most frequent SV type. The remaining
140 CGR-associated events were classified as **CGR-TRA** when HLAMP borders were formed by
141 interchromosomal translocations, and as **CGR-NOS** when no boundary translocations were
142 detected.

143

144 HLAMP Classifier (`eicicle`)

145 In this section, we describe `eicicle`, the algorithm we used to classify HLAMPs as either
146 extrachromosomal (ecDNA) or intrachromosomal (ICamp). It is based on the observation that the
147 copy number consistent with asymmetrical segregation of ecDNAs follows a broad distribution,
148 while the symmetric segregation of ICamps produces a more peaked pattern with copy numbers
149 clustered around a limited number of modes. For computational expediency, we first apply a Log-
150 transformation to the copy number data. In the Log-space, a Gaussian is equivalent to a Log-
151 Normal distribution capable of handling wide data. Our method is based on a structured Gaussian
152 mixture model with two one-dimensional components, where the mixture weights and the mean
153 and standard deviation parameters of each component are parameters of interest. To test a given
154 locus in a given case, our method takes as input the copy number values of all cells at that locus,
155 and outputs the estimated model parameters. We then aggregate these parameters to compute two

summary statistics per case: (a) an ecDNA score that is higher for ecDNA cases, and lower for ICamp cases, and (b) a mass-in-window score that is higher in ICamp cases and lower in ecDNA cases. Finally, a classifier based on Linear Discriminant Analysis (LDA) is trained on those cases with the most extreme values (top and bottom 5% quantile for each score) to find a decision boundary which is applied to the remaining cases. In the following, we set up the notation, describe the model, and details of inference.

Notation Let $x_1, \dots, x_N \in \mathbb{R}_{\geq 0}$ denote the observed data. We consider a mixture of $M = 2$ univariate Gaussians with means $\mu_0, \mu_1 \in \mathbb{R}$, standard deviations $\sigma_0, \sigma_1 > 0$, and mixing proportions $\pi = (\pi_0, \pi_1)$ with $\pi_k \geq 0$ and $\pi_0 + \pi_1 = 1$.

Model We consider two scenarios: either the distribution of the observed copy numbers is explained by a wide component, or by two well-separated narrow components. The former corresponds to a majority-ecDNA distribution, while the latter to a majority-ICamp distribution. To mitigate identifiability issues, we introduce a few constraints. To identify the means, we define μ_1 as a function of μ_0. To identify the component responsible for the ecDNA, without loss of generality, we assume that if a wide component exists, it is always the first one. To enforce this, we condition the second standard deviation σ_1 on the first one σ_0: if $\sigma_0 < \tau$ then we constrain $\sigma_1 \leq \sigma_0$, otherwise $\sigma_1 = 0.1$. This creates two distinct regimes: one where two narrow peaks explain the data, and another, where a mostly wide peak explains the data.

Generative Model The probabilistic generative model for `eicicle` is as follows:

$$s_0 \sim \mathrm{Beta}(\mathrm{a}_0, \mathrm{b}_0),$$

$$\sigma_0 \leftarrow g(s_0;\ e_0, e_1, e_0),$$

$$\mu_0 \sim \mathrm{Uniform}(0, x_{\max}),$$

$$\delta \sim \mathrm{Uniform}(d_{\min}, 0.5 x_{\max}),$$

$$\mu_1 \leftarrow \mu_0 + \delta,$$

$$s_1 \sim \mathrm{Uniform}(0,1),$$

$$\sigma_1 = \begin{cases} g(s_1;\ \sigma_0, e_1, e_2), & \sigma_0 < \tau \\ g(s_1;\ 0.1, e_3, e_4), & \sigma_0 \geq \tau \end{cases}$$

$$\pi = \begin{cases} \mathrm{Dirichlet}(\kappa_1, \kappa_1), & \sigma_0 < \tau \\ \mathrm{Dirichlet}(\kappa_2, \kappa_3), & \sigma_0 \geq \tau \end{cases}$$

$$x_n \mid \mu_{0:1}, \sigma_{0:1}, \pi \sim \sum_{m=0}^{1} \pi_m\ \mathcal{N}(x_n; \mu_m, \sigma_m^2).$$

We implemented this model in the probabilistic programming language Pyro[62] and used stochastic variational inference (SVI)[63] for inference. SVI is sensitive to initialization, so we repeated model fitting using five random seeds and selected the run with the lowest variational objective. In our implementation, we set $x_{\max} = \max_n x_n$, $\mathrm{a}_0 = \mathrm{b}_0 = 0.5$, $\tau = 0.3$, $d_{\min} = \sigma_{eff}/2.5$, where σ_{eff}

is defined as the robust spread of the data, computed as the difference between the 95th and 5th percentiles, ($\sigma_{eff} = Q_{0.95}(x) - Q_{0.05}(x)$). $\kappa_1 = 5$ when $\sigma_0 < \tau$ and $\kappa_2 = 1.0$ and $\kappa_3 = 0.5$ when $\sigma_0 \geq \tau$. We define a clamp function $g(z; s, \ell, u) = \min(\max(sz, \ell), u)$ and set $e_0 = x_{\max}/6$, $e_1 = 10^{-3}$, $e_2 = 2$, $e_3 = 10^{-4}$, $e_4 = 1$.

To summarize each case, we defined two scores: an ecDNA score ψ and a Mass-In-Window metric η. ψ is defined as a function of the mean and standard deviation of the Gaussian component with the largest π_k. That is: $\psi = (\pi_{\max})^2 q \overline{\sigma}_{\pi_{\max}} + (1-q)\overline{\mu}_{\pi_{\max}}$ where $\overline{\mu}$ and $\overline{\sigma}$ are the standardized mean and standard deviation respectively (across cases), and $q = 0.66$. η is defined as the mass in a small neighborhood of size $\epsilon = 0.5$ around the mode, that is $\eta = \sum_{m=0}^{1} \pi_m \int_{x_{\text{mode}}-\epsilon}^{x_{\text{mode}}+\epsilon} \mathcal{N}(x; \mu_m, \sigma_m^2)\mathrm{d}x$.

To classify cases into ecDNA vs. ICamp, we use linear discriminant analysis (LDA) using the 5% of extreme cases for training, as implemented in `discriminant_analysis.LinearDiscriminantAnalysis` function of `sklearn` library with default parameters. Probability bands at 0.05 and 0.95 are calculated using Platt scaling.

<u>ecDNA Simulator</u>

To investigate the dynamics of copy number alterations under ecDNA segregation, we developed a stochastic model. The simulation tracks a population of cells across discrete generations, where the state of each cell i at generation t is its integer copy number $X_{t,i} \in \mathbb{N}$. Let N_t denote the population size at generation t, and let K denote the imposed maximum population size (carrying capacity).

Population dynamics Population growth is implemented through a density dependent division probability $\lambda(N_t)$ that decreases as the population approaches capacity. In the log-scaled density-dependent division mode used here, each surviving cell divides independently with probability

$$\lambda(N_t) = \lambda_0 \left(1 - \frac{\log(N_t + 1)}{\log(K + 1)}\right),$$

where λ_0 is the base division probability.

Selection is modeled via a copy number dependent death probability $\mu(x)$ that induces a U-shaped fitness landscape. Specifically, we apply piecewise constant multipliers to a base death probability μ_0:

$$\mu(x) = \mu_0\, m(x), \qquad m(x) = \begin{cases} m_{\text{low}}, & x < \ell, \\ m_{\text{mid}}, & \ell \leq x \leq h, \\ m_{\text{high}}, & x > h, \end{cases}$$

where $\ell$and h are lower and upper copy number thresholds and $(m_{\text{low}}, m_{\text{mid}}, m_{\text{high}})$ are scenario specific multipliers. In addition, cells with $x < 2$ are deterministically removed.

Copy number dynamics Each generation consists of three stochastic steps applied to the current population:

1. **Death step.** For each cell with copy number x, we sample death as a Bernoulli event with probability $\mu(x)$. Cells that die or have $x < 2$ are removed from the population.
2. **Division decision.** Each surviving cell divides as a Bernoulli event with probability $\lambda(N_t)$.
3. **Copy number update on division.** For each dividing cell with copy number x, we first duplicate the copy number ($x \to 2x$) to mimic DNA replication. These duplicated copies are then partitioned between two daughters via a binomial split:

$$Y \sim \text{Binomial}(2x, p),$$

$$X_{t+1,\text{d1}} = Y,$$

$$X_{t+1,\text{d2}} = 2x - Y,$$

where $p \in [0,1]$ controls segregation bias ($p = 0.5$ is unbiased; $p > 0.5$ biases inheritance toward daughter 1). The two daughters are added to the next generation population. This procedure defines a discrete time, density-regulated branching process with copy number dependent selection and stochastic segregation driven copy number variation. Symmetric division is implemented as a deterministic segregation of DNA material equally between the two daughter cells. Under this condition, the copy number of a clonal population does not change. In our simulations, we used $K = 10^8$, $\lambda_0 = 1.0$, $\mu_0 = 0.01$with thresholds $\ell = 100$ and $h = 500$ and multipliers $m_{\text{low}} = 2.0$, $m_{\text{high}} = 4.0$, enforced $x \geq 2$ viability, followed by $p = 0.6$. $m_{\text{mid}} = 0.001$ and $m_{\text{mid}} = 10.0$ for high- and low-fitness scenarios respectively.

Extrachromosomal Amplicon De-Mixer (`ECADeMix`)

Given a single-cell WGS dataset in which specific regions have been identified as likely ecDNA, our goal is to infer a) the copy number profiles of distinct ecDNA species present (i.e., *amplicons*) and b) the multiplicity of each species in each cell (i.e., amplicon *loadings*). To solve this problem, we developed a coordinate ascent algorithm using a mixed integer linear program (MILP) to alternatively fix one of these quantities and infer the other, starting by fixing the initial amplicons.

The input to `ECADeMix` is single-cell CN profiles in 10kb bins, limited to specific regions of interest. To obtain this input, we normalized 10kb read counts using the constant of proportionality between read counts and copy number (i.e., the number of reads per copy per base) from 500kb-based HMMCopy results. Specifically, we took the average number of reads per copy in 500kb bins, divided this number by 50 (to account for the difference in bin size), then divided the 10kb read counts by this value to obtain approximate 10kb copy numbers. We normalized these profiles using non-amplicon-containing cells from the same sample (when available) to account for bin-specific read count biases.

$$\min_{A,Y} \quad \sum_{i=1}^{m}\sum_{j=1}^{n}\left|D_{i,j} - \sum_{p=1}^{k} Y_{i,p}A_{p,j}\right| + \lambda_A\left||A|\right|_0 + \lambda_Y\left||Y|\right|_0 + \lambda_{Y1}\left||Y|\right|_1$$

s.t.

$$A_{p,j} \in \{0,\ 1,\ \ldots z\}$$

$$Y_{i,p} \geq 0$$

for given number n of cells, number m of bins, number k of amplicons, maximum amplicon copy number z, and regularization parameters λ_A, λ_Y, and λ_{Y1}.

To initialize the MILP, we used either non-negative matrix factorization (scikit-learn implementation) or manual analysis of single-cell CN profiles to generate an initial set of amplicons. The number of amplicons, maximum amplicon CN, and regularization parameters were tuned on a case-by-case basis by inspecting residuals to select the minimal values of k and z that sufficiently explain the observed data. The parameters (and initialization files where relevant) used for each patient are available in the `ECADeMix` section of the paper repository (`hlamp/pipelines/ECADeMix/metadata`). We solve this MILP using the Gurobi optimizer.

<u>Calling CN Peaks</u>

To identify CN peaks, i.e., regions where cells cluster around a copy number mode, we used a histogram-based approach followed by 1D smoothing and local-maximum calling. Peaks were defined as local maxima in a one-dimensional signal, subject to user-specified constraints on peak height, prominence, separation, and width. For each locus-specific vector of per-cell copy-number values, we first constructed a 1D histogram (100 bins). The resulting bin counts were smoothed with a Gaussian kernel (`scipy.ndimage.gaussian_filter1d`, $\sigma = 1.5$), and peaks were identified as local maxima in the smoothed histogram using the `find_peaks` function (`scipy.signal`) with minimum peak height 1.5, minimum prominence 1.5, minimum peak distance 1 bin, and minimum peak width of 2 bins.

To test whether clones were preferentially enriched in a single peak (rather than distributed across multiple peaks), we computed a clone concentration metric per HLAMP region. For each clone, we computed the fraction of its cells in its most frequent peak and averaged this quantity over clones. For each case, we randomly permuted peak labels across cells within the dataset while preserving clone sizes and peak totals and recomputing the concentration metric to form a null distribution (n = 3,000 permutations). Finally, we combined the one-sided p-values across datasets using Stouffer's inverse-normal method (as implemented in `scipy.stats.combine_pvalues, method="stouffer"`), yielding a single global significance estimate for preferential single-peak enrichment. Prior to Stouffer transformation, we clipped per-dataset permutation p-values away from 0 and 1 to avoid infinite Z-scores.

295

296 Subclone CN Comparisons

297 For each ICamp region, we compared subclone-level copy-number patterns across three intervals:
298 i) the full amplicon, ii) the peak (maximum copy) segment within the amplicon, and iii) the interval
299 from the amplicon to the centromere on the same chromosome arm. Using signals package, we
300 computed subclone-level consensus copy number for each interval. Differences in consensus copy
301 number between subclones were evaluated by Student's *t* test on high-quality cancer cells.

302 To investigate the mechanisms of ICamp modulation between subclones, we quantified copy-
303 number similarity between subclones by computing the cosine similarity between their copy-
304 number profiles and derived a copy-number similarity score, based on the cosine similarity of the
305 subclone with lower weighted mean relative to the copy-number difference vector between the
306 two subclones.

307 To differentiate numeric vs. structural modulation of ICamp, we applied a Gaussian mixture model
308 with two components to the copy number similarity score (using `GaussianMixture` in
309 `sklearn.mixture`). Datapoints with predicted probability over 0.95 where annotated with their
310 respective class labels: those belonging to the class with the smaller mean CN similarity score
311 were designated as Numeric and the other as Structural. Datapoints with predicted probability less
312 than 0.95 are annotated as Intermediate.

313

314 Copy Number Correlation between Segments

315 Copy number of segments that belong to ecDNA were pairwisely compared using Pearson's
316 correlation, after removing cells with top and bottom 2.5% of the copy-number values to stabilize
317 the correlation. Then, hierarchical clustering was applied to identify the segments where the copy
318 number co-varies. The result was plotted in heatmap using `ComplexHeatmap` package in R
319 (https://jokergoo.github.io/ComplexHeatmap/).

320

321 Amplicon Architect

322 We ran AmpliconSuite-pipeline v1.3.5[64] for all available scWGS data for each library. We
323 complied result table from Amplicon Classifier function in each case to assess suggested
324 mechanisms of amplification. For case-level analysis, we considered ecDNA-positive when any
325 of the amplicon (from any library) was classified as ecDNA. For region-based analysis, we
326 annotated our scWGS-based HLAMP regions with Amplicon Classifier results, based on their four
327 mechanistic categories (linear, BFB, complex-non-cyclic, ecDNA). To quantify the agreement
328 between Amplicon Architect-based method and single cell copy-number distribution-based
329 classification, we calculated Cohen's kappa using `psych` package in R.

330 **scRNA-seq Analyses**

331 Preprocessing

All analyses for single cell RNA sequencing data were performed in Python (v3.10.16) using `Scanpy` (v1.10.4)[65]. FASTQ sequencing files were aligned to the GRCh38 reference genome, barcodes were filtered, and unique molecular identifiers (UMIs) were identified using the 10x `Cell Ranger` software (v6.0.1). Cells with fewer than 3 genes or with more than 40% mitochondrial content were removed. Highly variable genes were selected using the `highly_variable_genes` function with the flavor option `seurat_v3`. Raw counts were then normalized by library size and log-transformed with a pseudo-count of 1. Principal component analysis (PCA) was performed retaining the top 50 PCs, which were used for Leiden clustering and for generating the UMAP[66] embedding for visualization.

Cell Type Annotation

The cell types were defined using the markers reported in the previous study by our group[67]. Each cell was scored for each of the main cell types using the function `score_genes`. Then resulting scores were z-score normalized, then averaged per leiden cluster. Each cluster was annotated based on the cell type with the highest score. In turn, each cell inherited its cluster's annotation.

Identifying Malignant Cells

To identify malignant cells in the epithelial compartment, single-cell copy number signal was computed per sample using `infercnpy` v0.6.0 (https://github.com/icbi-lab/infercnvpy)[68], using a subset of non-epithelial cells as reference. Copy number clusters were identified by first computing top Principal components, and then community-based clustering.

Pathway Analysis

CCND1 loss in OV-052 Differential expression was performed for copy number cluster cluster "3" versus all other cells using `Scanpy`'s Wilcoxon rank sum test using function `rank_genes_groups` after library size normalization and log transformation (and identification of highly variable genes). Preranked gene list was constructed by sorting genes by signed effect size (log fold change). This ranked list was used as input to `GSEApy` (v1.1.4)'s preranked GSEA (`prerank`) against the MSigDB Hallmark gene sets with 1,000 permutations and gene set sizes restricted to 15 to 500 genes. Enrichment results were saved and summarized with a bar plot of the top positively and negatively enriched pathways.

Pathway activation in GBM0721 Gene sets for the Hallmarks pathway were acquired from msigdb_v2024[68]. For each hallmark pathway score, a multivariable linear regression model was fit with z scored EGFR and MDM4 expression as predictors, including standard scRNA seq QC covariates (total counts, detected genes, mitochondrial fraction) and an interaction term. Specifically, for each score y, the model $y \sim x_1 + x_2 + x_1 x_2 + \mathrm{QC}$ was estimated by ordinary least squares after removing cells with missing values, and scores with fewer than 10 complete observations were excluded. Benjamini–Hochberg correction was applied across pathway scores

370 separately for each coefficient. Pathways were flagged as showing a concordant joint association
371 when both x_1 and x_2 were significant after FDR correction ($q \leq 0.05$) and their coefficients had
372 the same sign.
373
374 Enrichment and Depletion of co-expression Patterns in scRNAseq

375 For each dataset and gene pair, we first defined per-cell "overexpression" using raw UMI counts
376 and a depth-adjusted Negative Binomial (NB) model. For each gene g in cell i, the expected mean
377 count μ_{ig} was set proportional to library size total_counts_i, and gene-specific dispersion θ_g was
378 estimated under an NB2 variance parameterization $\text{Var}(y_{ig}) = \mu_{ig} + \mu_{ig}^2/\theta_g$. We then computed
379 an NB Pearson residual

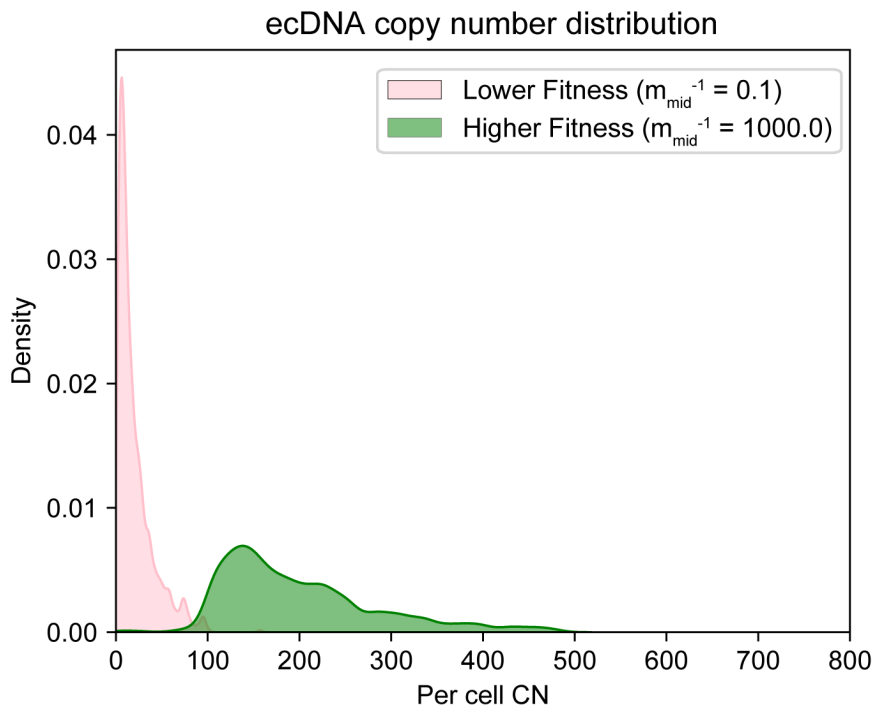
381 $$r_{ig} = \frac{y_{ig} - \mu_{ig}}{\sqrt{\mu_{ig} + \mu_{ig}^2/\theta_g}},$$

380

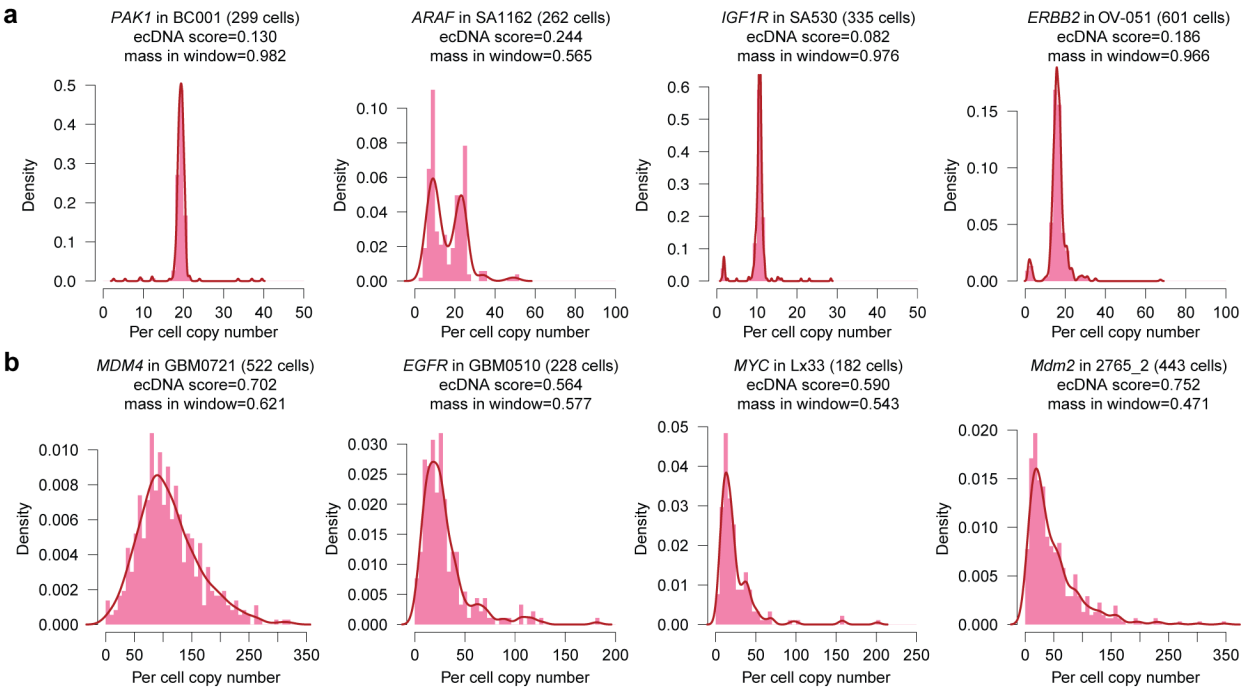
382 and called a gene "high" in a cell if $r_{ig} > 2$. Cells were assigned to one of four categories for each
383 pair (EGFR-only high, partner-only high, both high, neither), yielding a 2×2contingency table
384 $\begin{pmatrix} \text{both} & \text{EGFR-only} \\ \text{partner-only} & \text{neither} \end{pmatrix}$. We tested for association between "EGFR-high" and "partner-high"
385 using Fisher's exact test, with a one-sided alternative for depletion (*EGFR–CDK4*) or enrichment
386 (*EGFR–MDM4*), and report odds ratios and p-values as above.
387
388 **Fluorescence in situ Hybridization**
389 Cell lines were cultured in their respective media and arrested in metaphase with KaryoMAX™
390 Colcemid™ Solution (Thermo Fisher Scientific, #15212012). Colcemid conditions were 0.05
391 µg/mL for 2 hours (NCI-H524, NCI-H69, NCI-H82), 0.05 µg/mL for 6 hours (COLO320-DM,
392 COLO320-HSR), 0.05 µg/mL for 6 hours to overnight (GBM39-DM, GBM39-HSR), or 0.10
393 µg/mL for 2 hours (2765_2). Cells were harvested as a single-cell suspension, incubated in 75 mM
394 KCl for 10 minutes at 37°C, and fixed in ice-cold methanol:glacial acetic acid (3:1; Carnoy's
395 fixative) at −20°C for 30 minutes to overnight. Cell pellets were washed 3 additional times with
396 Carnoy's fixative and dropped onto glass slides to generate metaphase spreads.
397 Slides were processed for fluorescence *in situ* hybridization (FISH) through graded ethanol (70%,
398 85%, and 100%) and air-dried. FISH probes were either purchased (e.g., *MYC*, MetaSystems #D-
399 6008-100-OG; *MYCN*, Empire Genomics SKU *MYCN*, CHR02) or synthesized in-house from
400 BAC clones (*Mdm2* RP23-428D5; *Cen10* RP23-309H16; *EGFR* RP11-433C10 and RP11-
401 339F13; *CEN7* p7t1). For COLO320-DM/COLO320-HSR lines, DNA paint probes
402 (MetaSystems) were applied to slides, sealed with a coverslip using rubber cement, co-denatured
403 with the specimen at 72°C for 3 minutes, and hybridized overnight at 37°C in a humidified
404 chamber. Post-hybridization, slides were washed in 2X SSC three times for 2 minutes each, rinsed

405 in PBS, counterstained with DAPI, dehydrated through graded ethanol (70%, 85%, and 100%),
406 and mounted in ProLong Gold antifade (all cells except COLO320 lines; #P36930) or ProLong
407 Diamond antifade (for COLO320-DM/COLO320-HSR; Thermo Fisher, #P36965).

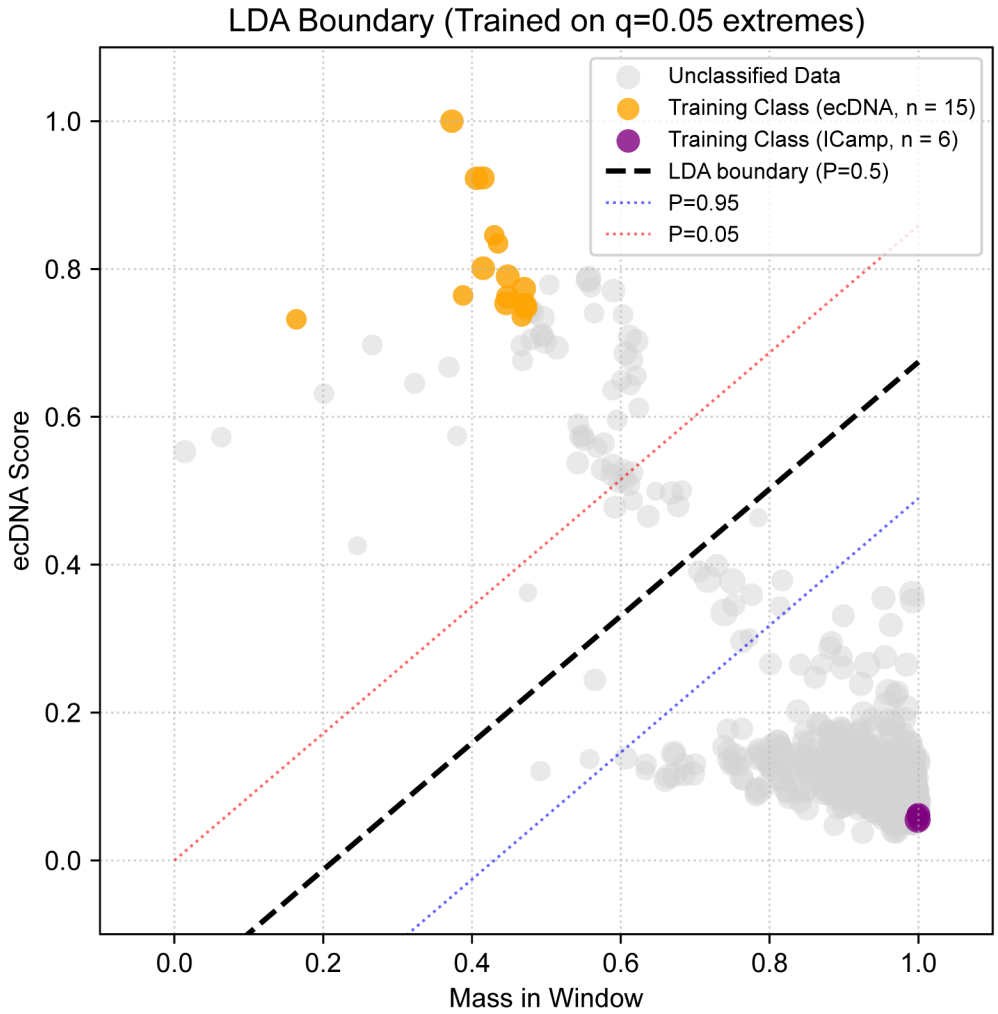
408 Images were acquired either on a Zeiss LSM880 confocal microscope (Carl Zeiss Microscopy)
409 using a Zeiss 63X/1.41 NA oil immersion objective (COLO320-DM/COLO320-HSR) or on an
410 ECHO Revolve microscope (Discover Echo) using a 60x oil objective (all other cell lines).



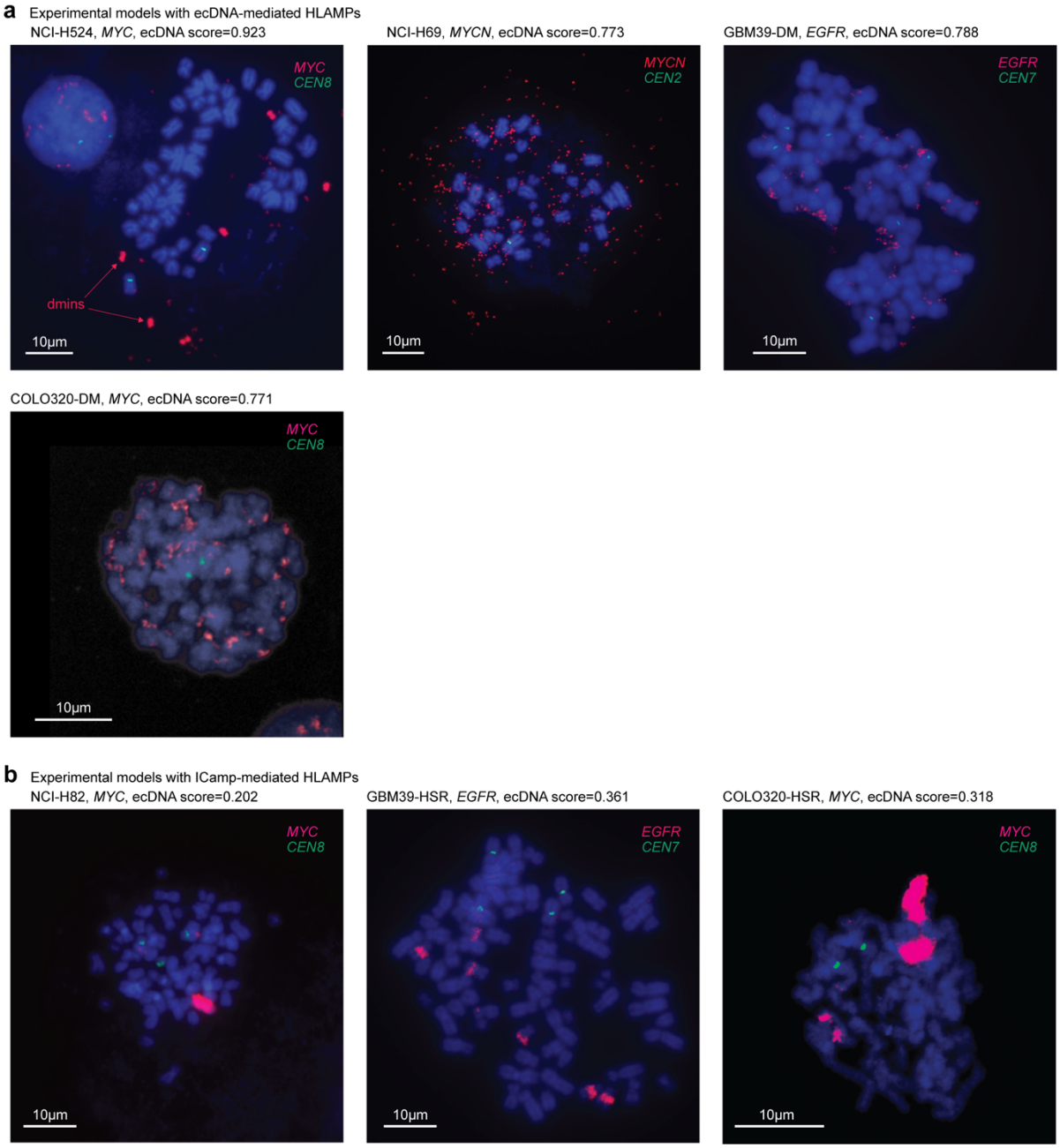
Extended Data Fig. 1. Probabilistic simulation model for high-level copy-number amplification. Copy-number histograms generated from simulations of ecDNA segregation after 500 generations; 1,000 cells were randomly subsampled to mimic sequencing. The green population has higher net fitness than the pink population. In this model, cells with copy number >500 incur a twofold higher fitness cost than cells with copy number <100. Net fitness for cells with copy number 100-500 is shown in the legend (**Methods**).



Extended Data Fig. 2. Distinct single-cell copy-number distributions for ecDNA and ICamp. **a-b**, Copy-number histograms (pink bars) and corresponding density estimates (red line) for HLAMP regions harboring *bona fide* oncogenes in 4 cases with ICamp **(a)** and 4 ecDNA cases **(b)**. Density plots were computed by kernel density estimation in base R.

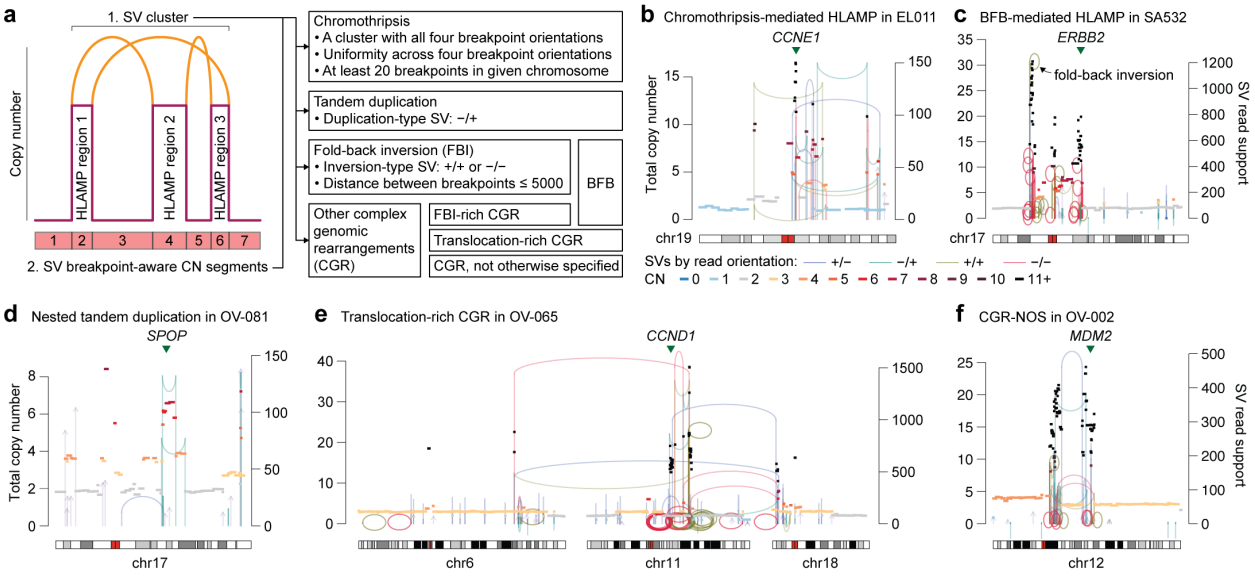


Extended Data Fig. 3. Machine learning-based differentiation of ecDNA from ICamp using `eicicle`. A probabilistic model scores each HLAMP region according to the properties of its single-cell copy number distribution. A linear classifier, trained on the extreme values of these scores, defines a decision boundary between ecDNA-positive and ecDNA-negative regions (**Methods**). Red and blue dotted lines show the 0.95 and 0.05 probability bands, respectively. LDA; linear discriminant analysis.

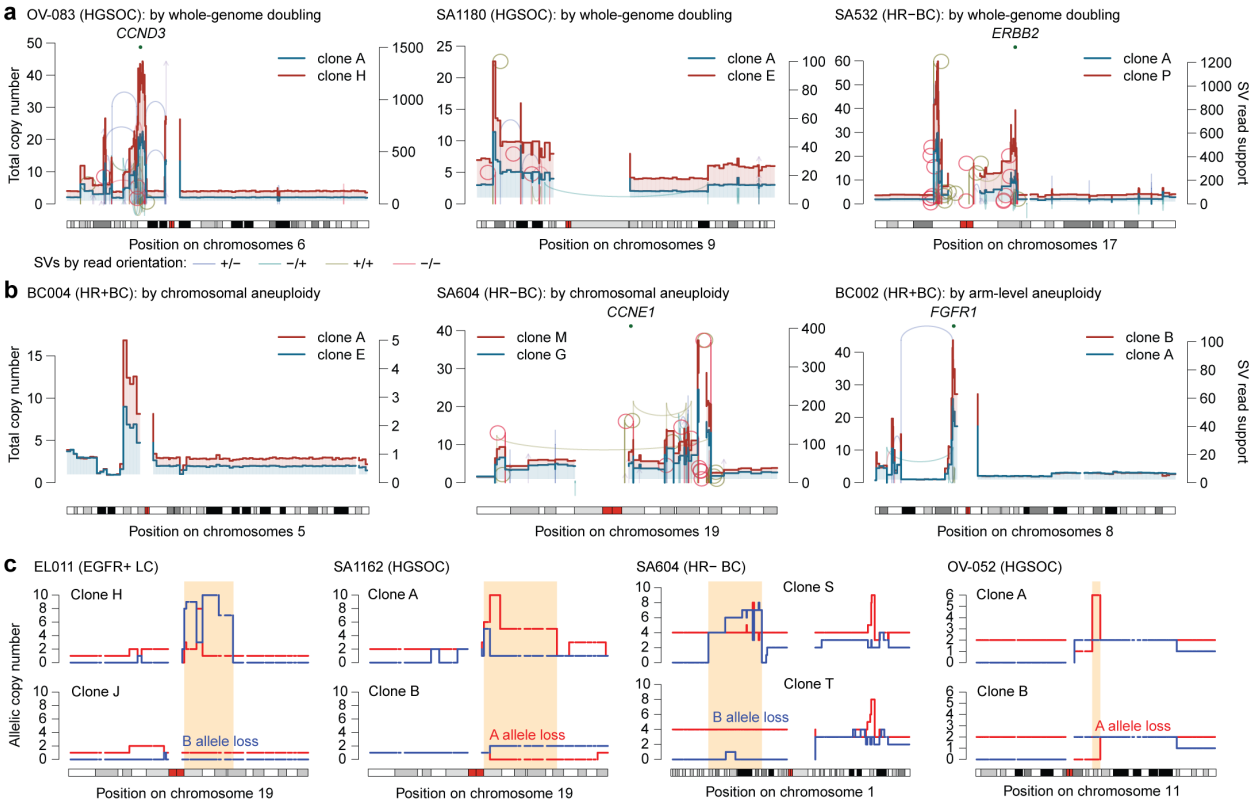


Extended Data Fig. 4. Cytogenetic features of HLAMP regions in experimental models. a-c, Cytogenetic evidence for oncogene amplification assessed by metaphase DNA fluorescence *in situ* hybridization (FISH). **a**, NCI-H524 (**top left**) is a small-cell lung cancer (SCLC) suspension cell line with high ecDNA score. The *MYC* probe (red) highlights a compact, intensely fluorescent extrachromosomal structure with bilobed morphology, consistent with double minutes (dmin). NCI-H69 (**top middle**) is a SCLC cell line grown in suspension. *MYCN* FISH shows numerous small, dispersed extrachromosomal signals consistent with *MYCN*-bearing extrachromosomal circular DNA (ecDNA). GBM39-DM (**top right**) and COLO320-DM (**bottom left**) show ecDNA-

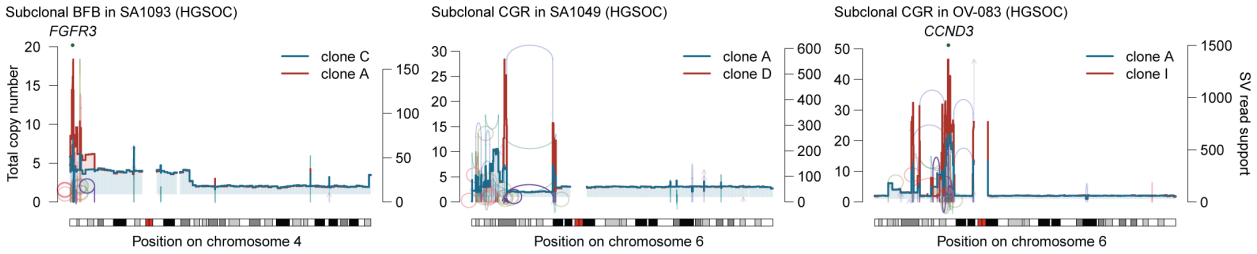
446 mediated amplification of *EGFR* and *MYC*, respectively. Consistent with these findings, these
447 models had high ecDNA scores in our single-cell copy-number (CN) distribution analysis using
448 `eicicle`. **b**, NCI-H82 (**left**) is an SCLC cell line with mix adherent and suspension growth. The
449 *MYC* probe reveals intrachromosomal clustered amplification as a homogeneously staining region
450 (HSR), with no ecDNA detected. GBM39-HSR (**middle**) and COLO320-HSR (**right**) show HSR-
451 mediated amplification of *EGFR* and *MYC*, respectively. Consistent with these findings, these
452 models had low ecDNA scores.
453



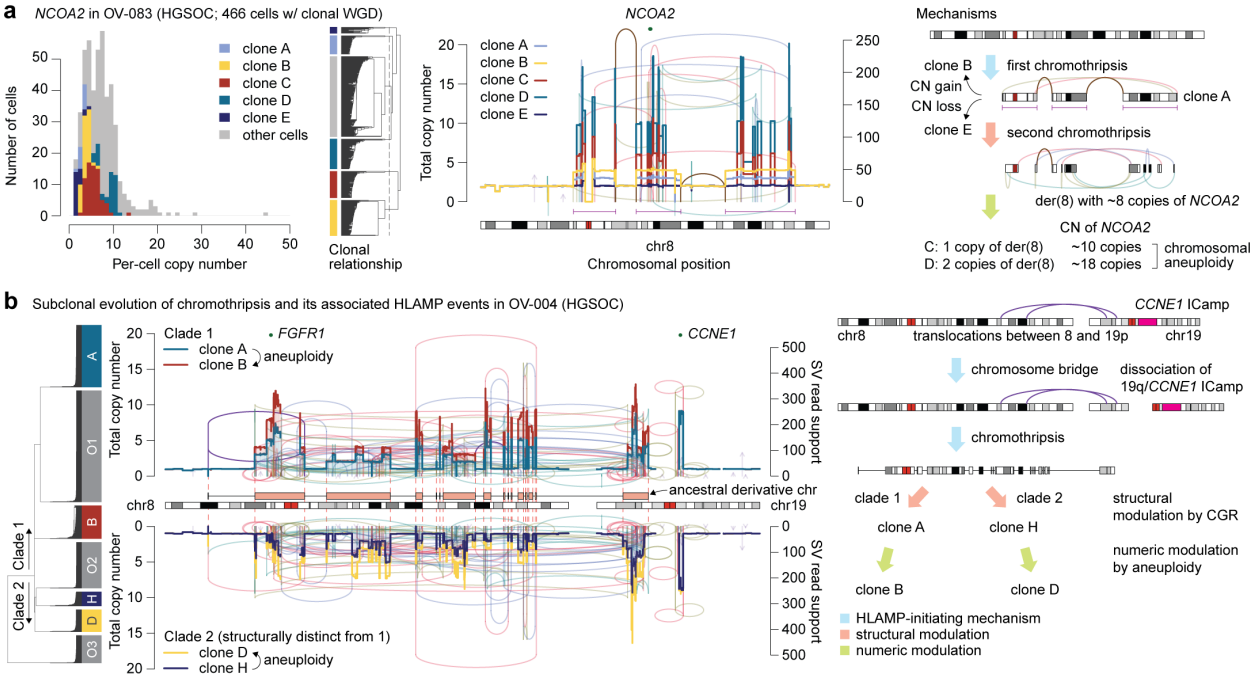
Extended Data Fig. 5. Classification of HLAMP regions by initiating mechanism. a, Schematic illustration of HLAMP regions and their relationship to structural variants (SVs). SV breakpoints were clustered into event footprints using `clusterSV`[23], accounting for breakpoint proximity, background SV rates and size distributions for each SV type. HLAMP regions were classified into 5 mechanistic categories (and unknown; **Methods**) based on SV features at their boundaries[21]. **b**, A *CCND1* HLAMP event with multiple SVs in all 4 orientations, indicating chromosomal fragmentation and random ligation, consistent with chromothripsis. **c**, A multifocal HLAMP event amplifying *ERBB2*. Major copy-number boundaries exhibit fold-back inversions (FBIs), a genomic footprint suggestive of ligation between sister chromatids. Multiple FBIs facing each other suggest generation by breakage-fusion-bridge (BFB) cycles. **d**, Two overlapping tandem duplications that together quadrupled a segment containing *SPOP* in an ovarian cancer case. **e**, A complex genomic rearrangement amplifying *CCND1*. Interchromosomal translocations at multiple borders between HLAMP segments and unamplified regions suggest that these translocations preceded copy-number amplification. **f**, A complex genomic rearrangement within chromosome 12, amplifying *MDM2*.



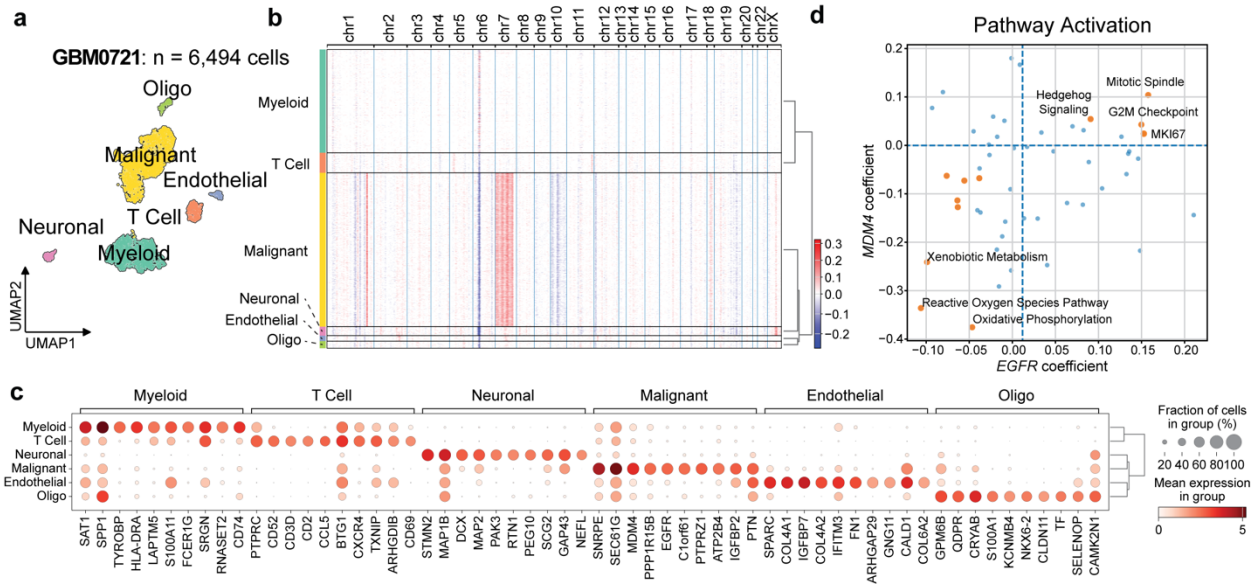
Extended Data Fig. 6. Numeric modulation of intrachromosomal amplification. a, Representative examples of subclone-specific duplication of ICamp via whole-genome doubling. In each case, the major clone (clone A; red line) exhibits a diploid-ranged copy number (CN) profile, whereas a minor subclone with whole-genome doubling (turquoise line) shows a global doubling of the CN profile, including the ICamp regions. **b**, Examples of aneuploidy-mediated ICamp duplication. ICamp is further amplified by subclone-specific whole-chromosomal aneuploidy (**left** and **middle** panels) and arm-level aneuploidy (**right** panel). **c**, Four cases showing subclone-specific loss of ICamp due to aneuploidy. Allele-specific CN analysis by the Signals package[17] reveals loss of the allele carrying the HLAMP event in subclones, resulting in loss of heterozygosity. These findings indicate that the subclone without HLAMP (exhibiting LOH) is not ancestral to the other subclones with HLAMP; instead, it represents a daughter lineage generated by aneuploid loss of the HLAMP-carrying allele. In EL001 and SA1162, the HLAMPs target *CCNE1*, and in OV-052 it targets *CCND1*.



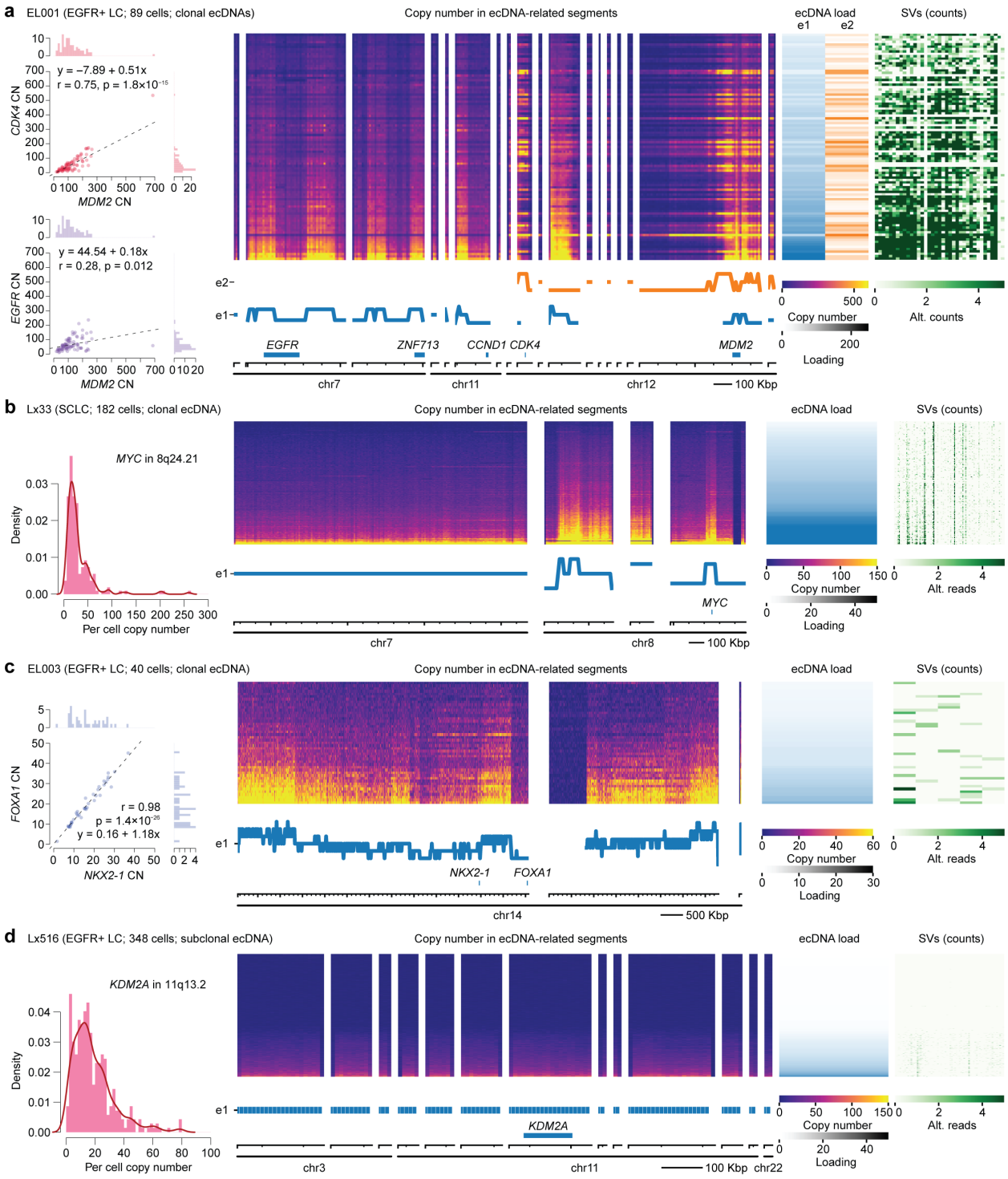
Extended Data Fig. 7. Structural modulation of intrachromosomal amplification. Structural modulation of HLAMP in three representative cases. A subclonal fold-back inversion (purple line) produces a subtelomeric copy-number (CN) gain in chromosome 4, further amplifying *FGFR3* (**left**). Subclonal complex genomic rearrangements (CGR; purple lines) increase CN at HLAMP regions (**middle** and **right**).



Extended Data Fig. 8. Diversification of HLAMP events after chromothripsis. a, An ICamp event involving chromosome 8 in an ovarian cancer case (OV-083), initiated by chromothripsis, followed by a second chromothripsis and subsequent whole-chromosome aneuploidy. **b**, Another ovarian cancer case (OV-004) with chromothripsis and subclone-specific CN profiles for two major phylogenetic clades (**left**). Both clades share the same segmental CN gain, indicating the ancestral structure of the derivative chromosome after chromothripsis (**middle**). This derivative chromosome underwent distinct repair trajectories, leading to divergence into two clades. Within each clade, subclones were further diversified by whole-chromosome duplication. A schematic illustration of the described mechanism (**right**).

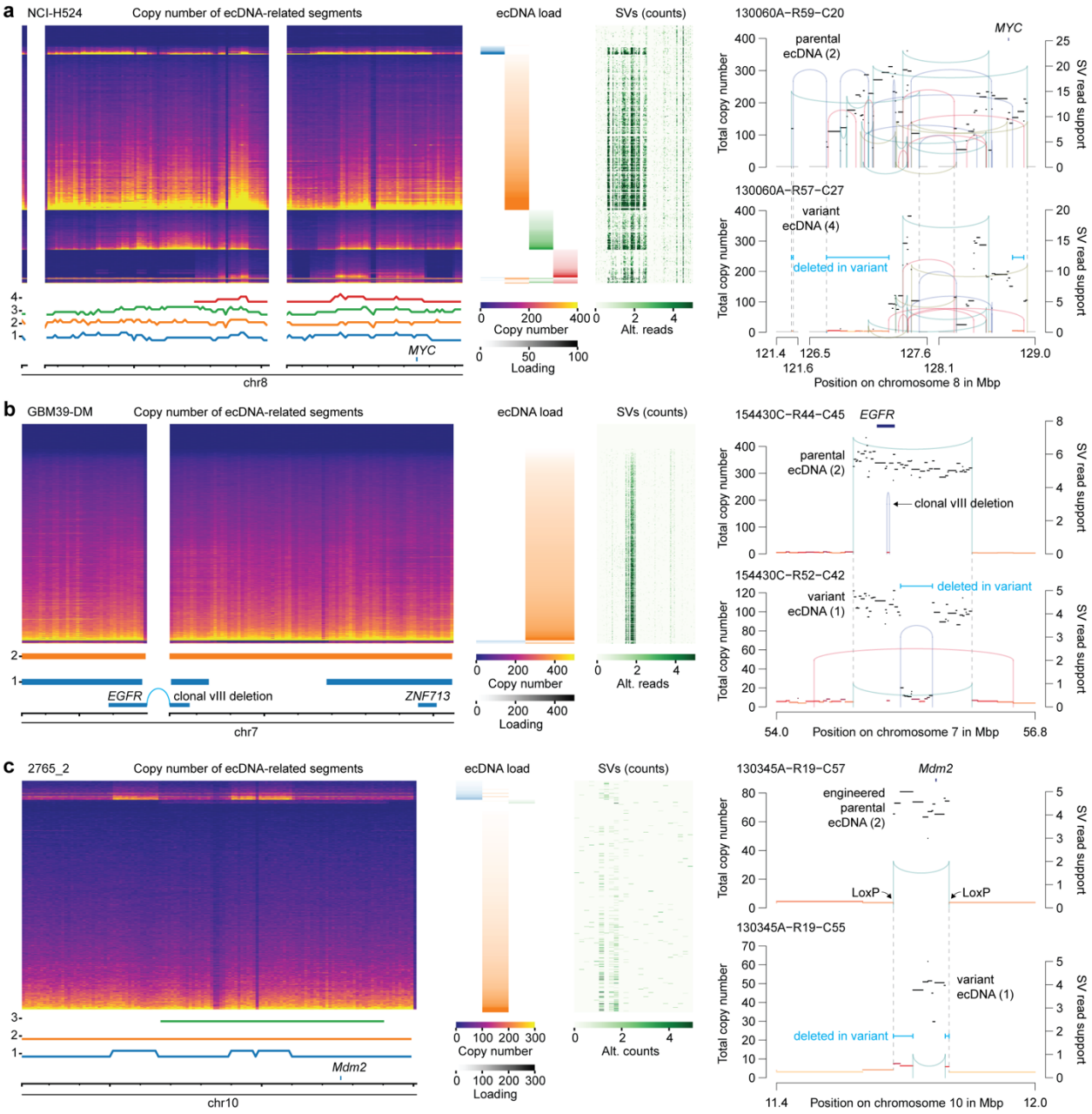


Extended Data Fig. 9. Single-cell RNA-seq analysis in glioblastoma samples. a-c, UMAP embedding of 6,494 cells passing quality control from scRNA sequencing of GBM0721 (**a**). Major cellular components are annotated using inferred copy number (CN) signals (**b**) and differentially expressed marker genes (**c**). **d**, Standardized partial regression coefficients linking hallmark pathway scores to expression of *EGFR* and *MDM4* in malignant cells, adjusted for quality metrics (**Methods**).

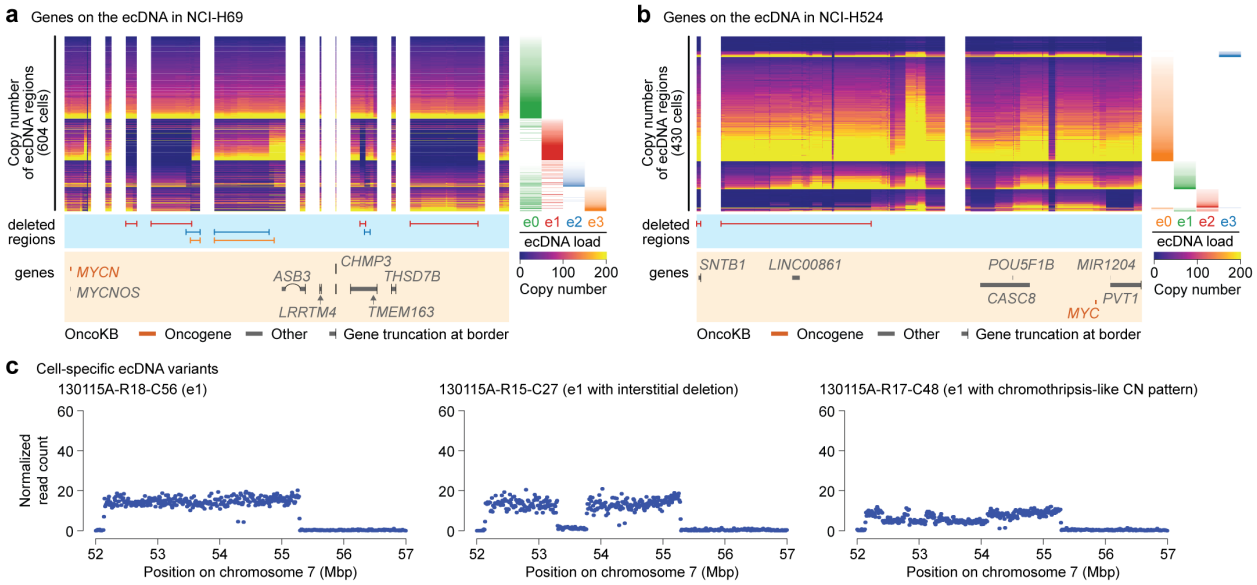


Extended Data Fig. 10. ecDNA architecture revealed by `ECADeMix`**.** Decomposition of ecDNA species based on `ECADeMix`. Shown are four cases in which ecDNA species exhibit minimal structural variation across cells **a-d**. Copy-number (CN) distribution of key oncogenes and their pairwise correlations (**left**), per-cell CN heatmaps and inferred structures of ecDNA species

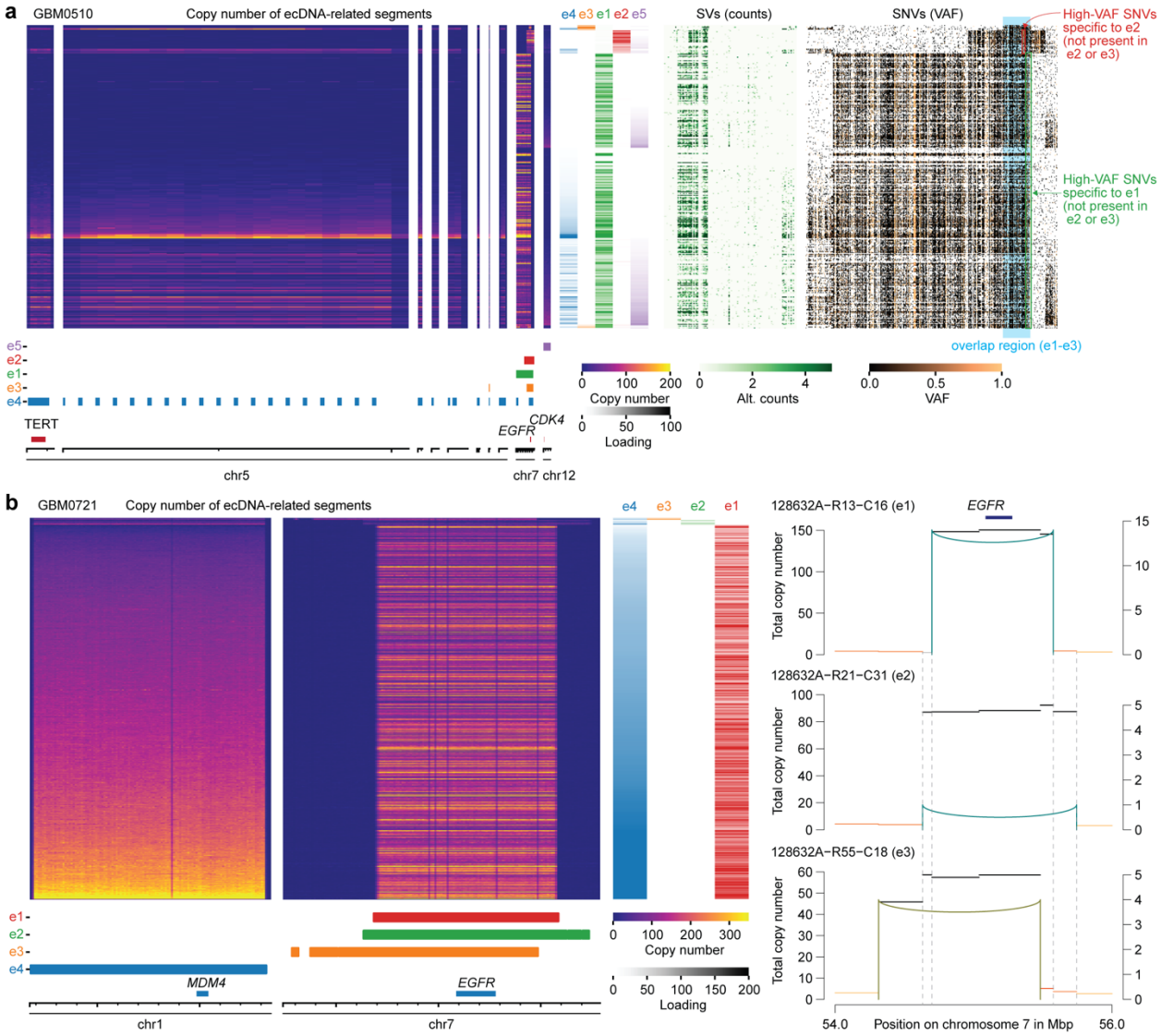
523 (**middle**), per-cell ecDNA species loads and associated SVs at single-cell resolution (**right**).
524 EL001 was sequenced at greater depth (≥0.1x per-cell coverage), increasing sensitivity for SV
525 detection.
526



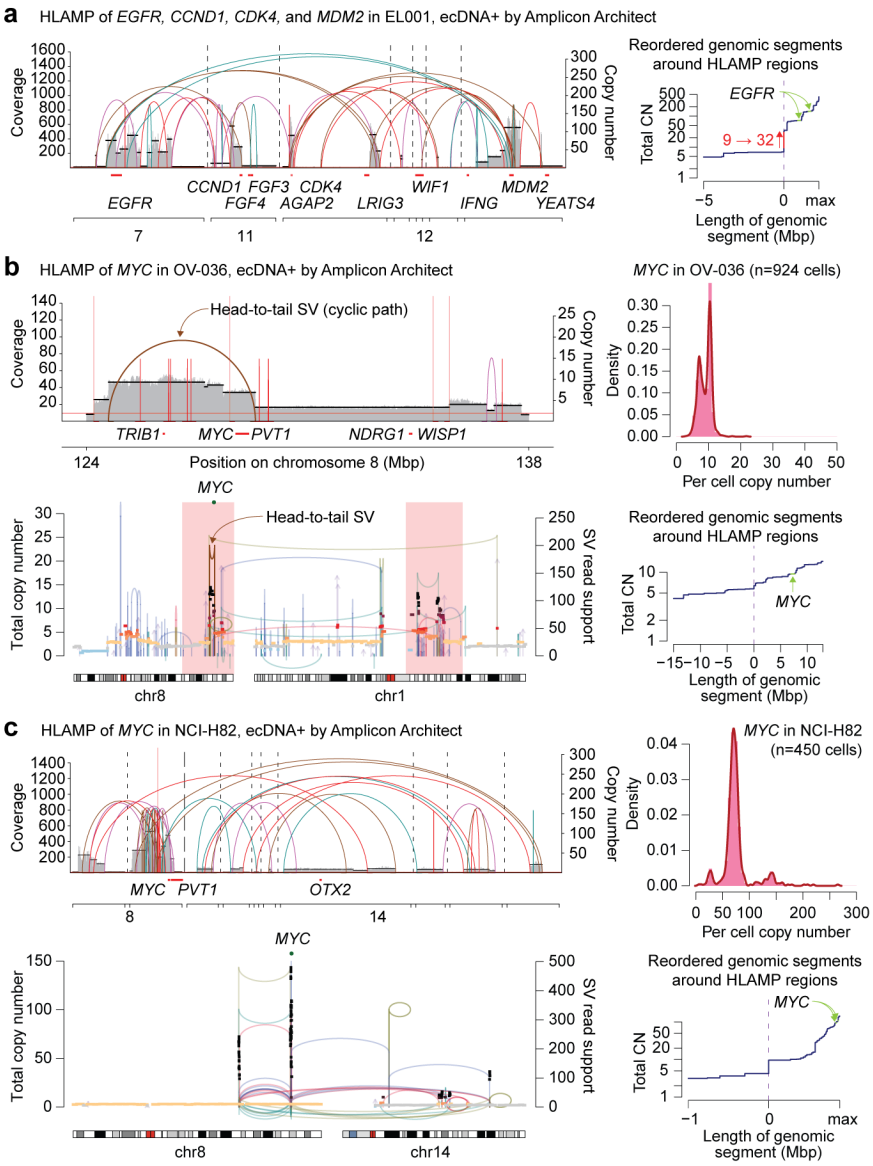
Extended Data Fig. 11. Divergent evolutionary trajectories driven by ecDNA remodeling. a-c, Three representative cases in which multiple ecDNA species target the same locus, with evidence of divergent evolution. In each case, one species spans the full set of HLAMP regions, consistent with the parental ecDNA structure (**left**). Variant ecDNA species harbor intramolecular rearrangements that produce segmental deletions (**right**). The SCLC cell line NCI-H524 shows ecDNA targeting *MYC* generated by complex genomic rearrangements **(a)**, GBM39-DM cells harbor *EGFR* ecDNA with the variant III deletion **(b)**, and the genetically engineered mouse model 2765_2 contains engineered ecDNA targeting *Mdm2* **(c)**.



Extended Data Fig. 12. Structural variants underlying divergent ecDNA evolution. a-b, Variant ecDNA-specific deletions and their relationships to genes on ecDNA molecules in NCI-H69 **(a)** and NCI-H524 **(b)**, two cases with prevalent ecDNA variants. In both cases, deletions primarily affect gene-poor regions, regions containing already-truncated genes, or long noncoding RNAs. In contrast, *bona fide* oncogenes (*MYCN* in NCI-H69 and *MYC* in NCI-H524) are preserved. **c**, Cell-specific copy-number (CN) patterns indicative of ongoing ecDNA evolution. Shown are an intact e1 ecDNA in GBM0510 (**left**), a variant with a central deletion (**middle**), and a variant with an oscillating CN pattern (**right**), consistent with chromothripsis on ecDNA.



Extended Data Fig. 13. Convergent evolution in glioblastoma via recurrent acquisition of *EGFR* ecDNAs. a, GBM0510 shows convergent evolution, with three *EGFR* ecDNA species arising from distinct structural variants (middle heatmap) and defining three mutually exclusive clones. These three clones share many single-nucleotide variants (SNVs) and differ by a small number of clone-specific SNVs (indicated by red and green arrows), consistent with late emergence of *EGFR* ecDNAs during tumor evolution. **b**, GBM0721 similarly harbors three distinct *EGFR* ecDNA species attributable to different structural variants (e1-e3). ecDNA e1 is predominant, whereas e2 and e3 are found in a minority of cells. In contrast, only a single *MDM4* ecDNA species is present, consistent with earlier acquisition. SVs underlying e3 could not be fully resolved because of the small number of e3-positive cells.



Extended Data Fig. 14. Discrepancies between bulk genome graph-based and single-cell copy number distribution-based ecDNA definitions. a, In EL001, Amplicon Architect (AA) classified the amplicon as ecDNA (**left**), consistent with its single-cell copy-number (CN) distribution. These regions exhibit an abrupt CN transition (from 9 to >30 copies; **right**), characteristic of ecDNA-positive cases. However, because of complex connectivity, AA grouped all segments into a single ecDNA, whereas our analysis resolved two independent ecDNAs (**Fig. 3**). **b**, In OV-036, AA identified a cyclic genome graph around the *MYC* locus (**top left**) and classified it as ecDNA. However, the single-cell CN distribution for *MYC* shows 2 prominent peaks (**top right**), consistent with an ICamp with intratumoral heterogeneity. At the pseudobulk level, the *MYC* amplicon is densely connected by complex genomic rearrangements to regions on chromosome 1 (**bottom left**). These segments (red shading) show a gradual, stepwise increase in CN toward the *MYC* locus (**bottom right**), a pattern more consistent with multistep amplification

576 than ecDNA formation. **c**, In NCI-H82, AA classified the *MYC* amplicon as ecDNA (**top left**). The
577 amplicon comprises complex genomic rearrangements in 8q and connections to 14q (**bottom left**).
578 Although the CN transition from non-amplified to amplified region shows a high junctional CN at
579 the pseudobulk level (**bottom right**), the single-cell CN distribution is consistent with ICamp (**top**
580 **right**). Metaphase fluorescence *in situ* hybridization (FISH) revealed an HSR and no evidence of
581 ecDNA (**Extended Data Fig. 4**), consistent with prior ecDNA amplification followed by
582 chromosomal integration.

583 **Method References**

584 57. Medina-Martínez, J. S. *et al.* Isabl Platform, a digital biobank for processing multimodal
585 patient data. *BMC Bioinformatics* **21**, 549 (2020).

586 58. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic
587 expression at nucleotide resolution reveals disrupted pathways in triple-negative breast
588 cancer. *Genome Res.* **22**, 1995–2007 (2012).

589 59. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural
590 variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 202
591 (2021).

592 60. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat.*
593 *Methods* **12**, 966–968 (2015).

594 61. Garsed, D. W. *et al.* The Architecture and Evolution of Cancer Neochromosomes. *Cancer*
595 *Cell* **26**, 653–667 (2014).

596 62. Bingham, E. *et al.* Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**,
597 973–978 (2019).

598 63. Hoffman, M., Blei, D. M., Wang, C. & Paisley, J. Stochastic Variational Inference. *arXiv*
599 *[stat.ML]* (2012) doi:10.48550/arXiv.1206.7051.

600 64. Luebeck, J. *et al.* AmpliconSuite: an end-to-end workflow for analyzing focal amplifications
601 in cancer genomes. *Bioinformatics* (2024).

602 65. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
603 analysis. *Genome Biol.* **19**, 15 (2018).

604 66. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
605 Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

606 67. Vázquez-García, I. *et al.* Ovarian cancer mutational processes drive site-specific immune
607 evasion. *Nature* **612**, 778–786 (2022).

608 68. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary
609 glioblastoma. *Science* **344**, 1396–1401 (2014).