nature methods

Article

https://doi.org/10.1038/s41592-026-03279-4

Echidna: a Bayesian framework for integrative inference of copy number alterations impacting phenotypic plasticity

Received: 15 December 2024

Accepted: 22 September 2026

Published online: 16 November 2026

Check for updates

Joy Linyue Fan[1,2,16], Mingxuan Zhang[2,3,4,16], William O'Brien[5,6], Joshua D. Myers[1,2], Sharanya Chatterjee[7], Kevin Hoffer-Hawlik[1,2], Nicolas Beltran-Velez[7], Alexander Preau[7], Isha Arora[2,8], Johannes C. Melms[9,10], Jana Biermann[9,10], Edridge D'Souza[9], Somnath Tagore[9], Timothy Olsen[11], Siran Li[12], Peter A. Sims[3,13], Benjamin Izar[9,10,14,15] ✉ & Elham Azizi[1,2,7,15] ✉

The ability of cells to adapt their behavior in response to genetic or environmental changes is a fundamental biological process that drives cellular diversity in both normal and pathological contexts, including in tumor evolution. Copy number alterations (CNAs), which can include large-scale changes in genomic ploidy, play a key role in creating adaptive advantages in evolving cell populations. However, tracing these populations across experiments and disentangling the role of CNAs from other sources of transcriptional variability remains challenging and expensive. Here we present Echidna, a Bayesian hierarchical model that integrates single-cell RNA sequencing and bulk whole-genome sequencing data to enable robust integrative inference of CNAs, jointly across samples and treatment time points, to better understand the impact of genomic events on cellular phenotype. Applying Echidna to patient tumor specimens, we demonstrate its superior performance in clonal reconstruction and derive insights into therapy resistance mechanisms.

The interplay between the genome and transcriptome governs cellular behavior in both health and disease, driving complex biological processes such as differentiation, adaptation and response to environmental changes[1–4]. An instructive example of this interplay is tumor evolution, where genomic instability and transcriptional variability lead to diverse phenotypic states, enabling cancer cells to adapt to selective pressures such as immune surveillance and therapeutic interventions[5–9]. However, the genome and transcriptome have largely been explored as disjoint features only.

Genomic analyses using whole-genome sequencing (WGS) have established that chromosomal instability and somatic copy number alterations (CNAs) drive tumor progression, including through the accumulation of mutations in oncogenes and loss of tumor suppressors, giving rise to genetically distinct clones with variable fitness.[10–18] In parallel, single-cell RNA sequencing (scRNA-seq) studies have focused on transcriptional variability, revealing extensive heterogeneity across phenotypic states[19–21], associated with disease progression or response to therapies[22,23].

This transcriptional diversity reflects a phenomenon known as phenotypic plasticity, wherein a clone transitions between states (Fig. 1a). In addition to genetic alterations, epigenetic or transcriptional variability may drive critical phenotypes, such as resistance to therapies[24]. It remains unclear to what extent transcriptional plasticity is driven by intrinsic genetic features (gene dosage effect (GDX)[25]), compared to extrinsic, microenvironmental stimuli.

This gap motivates the development of computational tools that can robustly estimate CNAs across multiple samples, while integrating genomic information with transcriptional phenotypes.

A full list of affiliations appears at the end of the paper. ✉e-mail: bi2175@cumc.columbia.edu; ea2690@columbia.edu

https://doi.org/10.1038/s41592-026-03279-4

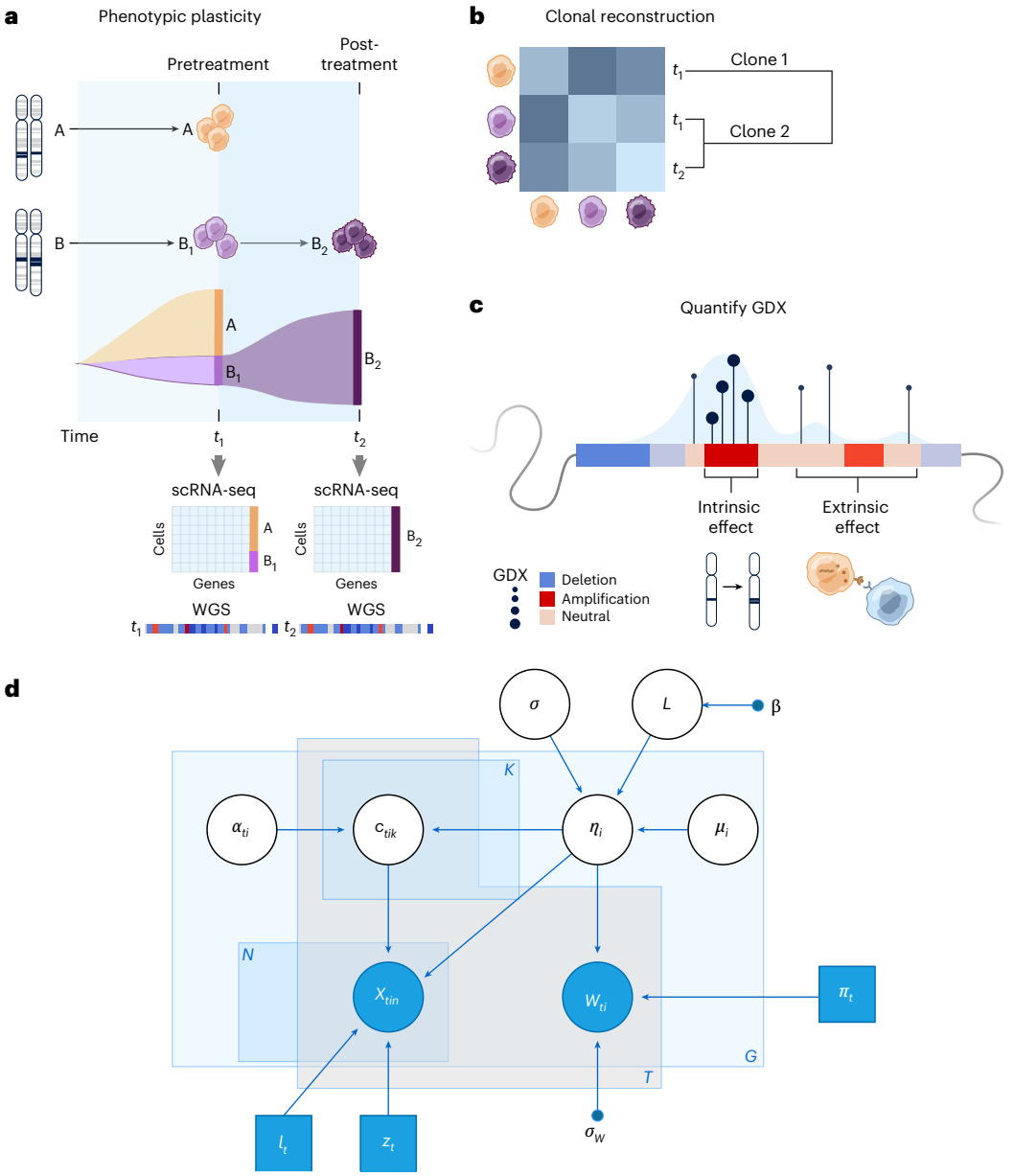


Fig. 1 | The Echidna model. a, Echidna integrates single or multiple time point scRNA-seq and WGS data to characterize the driving forces of phenotypic plasticity and clonal expansion. **b**, Echidna infers cluster-level gene dosage values, while also learning the correlation between clusters to reconstruct the underlying clonal structure. **c**, Echidna's hierarchical structure allows for the computation of a novel metric, the GDX, that quantifies the extent to which a phenotypic shift in gene expression (shown in bars) may be attributed to a genetic alteration. The outputs of Echidna may be used to identify hotspots of DEGs, while identifying their dependence on genetic alterations through GDX (shown with dot size). **d**, Plate model of Echidna. White nodes represent latent random variables and shaded nodes represent observed random variables. Shaded squares represent parameters we fix according to observed data. Arrows indicate conditional dependency relationships and the direction represents the generative process. Plates indicate dimensions of independence. Credit: images in **a**–**c**, Sigrid Knemeyer/SciStories.

Previous attempts to quantify the link between the genome and transcriptome[26,27] relied on bulk sequencing, which obscures tumor heterogeneity and the contribution of noncancerous cells in the microenvironment. While paired genomic and transcriptional profiling at the single-cell level is technically possible[28–31], such data are difficult to generate at scale and depth, and datasets coupling these methods remain scarce[32,33]. Conversely, scRNA-seq analysis pipelines for clustering single-cell transcriptomics and identifying differentially expressed genes (DEGs) are also limited in the scope of biological insights as they merely focus on phenotypic plasticity and do not elucidate driving factors such as gene dosage, clonal identities, and lineages.

Previous methods for inferring single-cell CNAs from scRNA-seq data[34–36] have leveraged hidden Markov models (HMMs) to infer copy number states along the chromosome. However, this sequential constraint also restricts the ability to model the relationship between CNAs and expression. Thus, these methods cannot jointly consider the downstream transcriptomic effects, nor capture clonal evolution across time points and tissues. To address these challenges, we present Echidna, a Bayesian hierarchical model that integrates the genome and transcriptome across time points and samples to improve CNA inference, quantify the dependence between transcription and CNAs, and study clonal dynamics. Echidna requires only scRNA-seq and bulk WGS, both of which are readily obtainable. We demonstrate that

Echidna improves inference of clone-specific CNAs without requiring a normal reference or strict sequential assumptions, while modeling temporal dynamics linking tumor evolution to phenotypic plasticity. We anticipate that insights from Echidna will enable both the discovery of novel genome alterations driving tumor plasticity and resistance to therapy, as well as the transcriptional mechanisms through which these alterations take effect.

Results

Overview of Echidna

Echidna is a Bayesian hierarchical model that integrates genomic and transcriptomic data to study how CNAs shape transcriptional variability, phenotypic plasticity and clonal evolution. It jointly models bulk WGS and scRNA-seq data across one or multiple time points.

Echidna takes as input phenotypic states defined as clusters of cells from scRNA-seq data. When longitudinal data are available, cells from all time points are jointly clustered to enable detection of evolving clones. The scRNA-seq data are represented as $X^{T\times N\times G}$ for T time points, N cells and G genes. We denote the observed proportion of clusters within a sample or time point as $\pi \in \mathbb{R}^{T\times K}$, with K being the total number of clusters. Bulk WGS-derived gene dosage is mapped to genes and represented as $W^{T\times G}$.

The generative process consists of two main components (Fig. 1a). First, we assume bulk gene dosage measurements from WGS are a weighted sum of latent cluster-specific gene dosage profiles (η), with the weights corresponding to the cluster proportions (π) for each time point (t). To capture clonal relationships, cluster-specific gene dosage profiles are modeled as correlated multivariate Gaussian variables centered at neutral copy number (2), with covariance Σ encoding similarity between clusters (Fig. 1b; Methods).

Second, single-cell transcript count is modeled as a Gamma-Poisson distribution to account for over dispersion[37,38]. To link the two components, we define transcriptional rate as a latent parameter (c). This parameter follows a Gamma distribution with a rate dependent on gene dosage. The observed expression is then modeled as a function of this rate along with cluster- and cell-specific effects (Fig. 1c,d; Methods).

Integrating transcriptomic and WGS data improves accuracy in learning CNAs compared to methods that use scRNA-seq data alone, as the bulk gene dosage measures set a constraint on the sum of CNAs across cells. After model fitting, CNA profiles are refined through post-processing. Learned gene dosage profiles (η) are first smoothed along the genome using a one-dimensional (1D) Gaussian kernel. Then, cluster-specific neutral values are learned by a Gaussian mixture model (GMM) (Extended Data Fig. 1a), allowing us to correct the estimated gene dosage per cluster. Last, a Gaussian HMM model is applied to infer discrete copy number states (Extended Data Fig. 1b,c). Hierarchical clustering can then be performed on either the inferred correlation between gene dosage values of clusters or the estimated posterior covariance directly to obtain a proxy for a phylogenetic tree between clusters. The tree is then used to define clones as groups of clusters with similar CNAs.

A key feature of Echidna is its ability to track clones and their phenotypic plasticity over time. To achieve this interpretability, while avoiding full covariance matrix inference, we parameterize the covariance Σ using a Cholesky decomposition, with off-diagonal elements sampled from an LKJ (Lewandowski–Kurowicka–Joe) distribution, which is a distribution over positive definite symmetric matrices with unit diagonals [39]. This ensures numerically stable inference of cluster dependencies, while allowing the adjustment of subclonal resolution (Methods).

Finally, to study the influence of gene dosage on expression, we define a metric for GDX, which is computed based on posterior distributions learned by the model (Fig. 1c, Extended Data Fig. 1d and Supplementary Note A).

Echidna outperforms other methods in clonal reconstruction

To assess Echidna's ability to infer clonal architecture, we applied it to matched single-nucleus (sn) hybrid DNA-seq and RNA-seq datasets with corresponding WGS collected from frozen endometrial cancer tissues[40] (Fig. 2a). We limit our analyses to patients 2 and 5 (ref. 40) as these tumor specimens involved substantial genetic and phenotypic heterogeneity[40]. Sequenced using a novel adaptation of the BAG technology[40], these data provide paired DNA–RNA measurements, enabling ground-truth validation. Specifically, we fit our model to the WGS and snRNA-seq, and reserved the snDNA-seq data for validation.

In tumor 2, clones defined by single-cell (sc)DNA data show that genetically similar clones can exhibit heterogeneous phenotypic states. For example, clone T2B corresponds to distinct clusters 7 and 8 in scRNA-seq data that display functionally different transcriptional signatures from each other. Differential gene expression analysis reveals that the T2B clone is enriched in genes related to TNF signaling (*CCL20*, *TIPARP* and *NR4A3*) compared to other clones, while cluster 8 is enriched in genes involved in epithelial–mesenchymal transition (*COL7A1*, *DACT2*, *FSTL3* and *LAMA1*) compared to cluster 7 (Fig. 2a). Because of these transcriptional differences, cluster correlation computed using expression data alone does not reflect a strong similarity between clusters from the same clone (Fig. 2b). By contrast, CNA-based similarity derived from scDNA data (Fig. 2c, left) reveals a drastically different structure (quantified by Frobenius distance). The integrative nature of Echidna's deconvolution successfully recovers cluster-level genetic similarities that more closely resemble the ground-truth clonal structure while retaining some variability within clones (Fig. 2c). For example, clone T2A is further split into subclones (C2, C3 and C4), reflecting its phenotypic plasticity.

Echidna learns a covariance matrix that more accurately approximates the clonal structure compared to Numbat and InferCNV, as quantified by the Frobenius distance (Fig. 2c and Extended Data Fig. 2a). This improvement is partly driven by the integration of WGS data, which serves as a genetic reference in cases where the scRNA-seq may be of lower depth. As noted in ref. 40, the coverage of the novel hybrid sequencing method is lower than standard sn-sequencing methods for DNA and RNA separately.

To systematically evaluate performance, we assessed the clonal reconstruction accuracy for state-of-the-art methods, namely inferCNV[41], CopyKAT[42], Numbat[36] and XClone[43], using ground-truth labels derived from scDNA data (Methods). Echidna clones were defined by thresholding the dendrogram of the cluster–cluster correlation matrix. We computed three standard clustering metrics to compare performances: purity, which measures the extent to which inferred clones contain only cells from ground-truth clones; adjusted mutual information (AMI), which quantifies the agreement between the inferred ground-truth clone labels while adjusting for chance; and adjusted rand index (ARI), which considers all pairs of cells, counting pairs that are assigned in the same or different clones, while adjusting for chance (Methods). We observed significant improvement with Echidna compared to other methods with all metrics ($P < 1 \times 10^{-10}$; Fig. 2e). We also conducted sensitivity analysis to verify the necessity of WGS information. High noise levels of WGS lead to progressively worse clone reconstruction performance (Supplementary Fig. 2).

To demonstrate robustness to small or noisy clusters, we dropped the cluster with the fewest number of cells in tumor 2, and refit Echidna using the same hyperparameters and post-processing steps. The model reproduced its ability to detect the clone with the greatest phenotypic plasticity (C1; Extended Data Fig. 2b), and again reconstructed cluster dependencies consistent with ground-truth clonal structure (Extended Data Fig. 2b,c). To further demonstrate the model's robustness across a wider operational landscape, we conducted a systematic sensitivity analysis on the two key hyperparameters (initialization of the variational posterior correlation and LKJ concentration) under progressive subsampling of the dataset. Echidna was first trained

https://doi.org/10.1038/s41592-026-03279-4

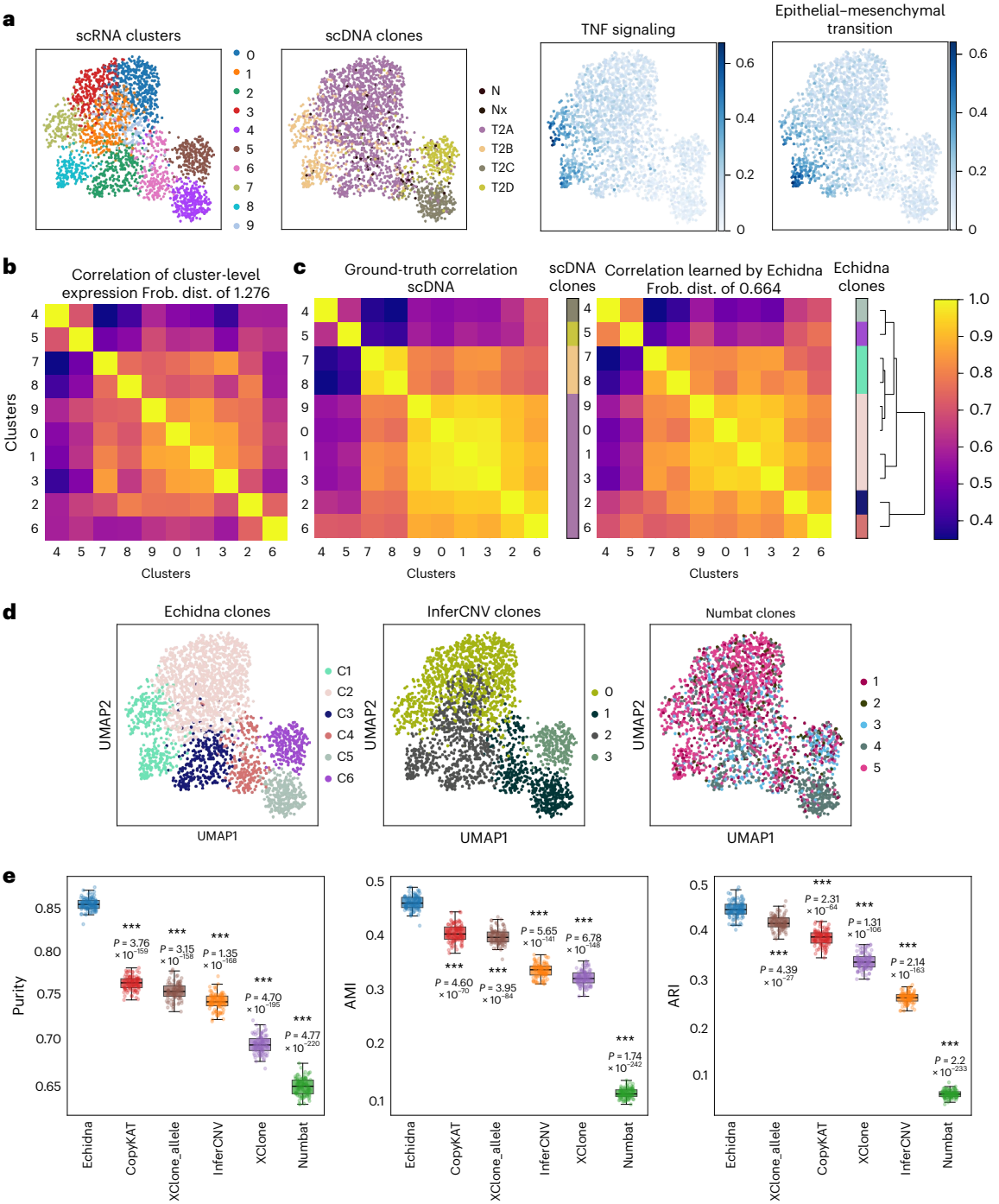


Fig. 2 | Echidna achieves accurate clone reconstruction by learning cluster-specific gene dosage and cluster correlation structure. a, UMAPs of the scRNA-seq data from specimens from patients with endometrial cancer[40] where each dot represents a cell. The cells are colored by Leiden clusters identified from expression (left), clones identified by matching scDNA data (middle) and expression of gene signatures enriched in clone T2B (right). **b**, Correlation matrix of pseudo-bulked scRNA expression at the cluster level. The Frobenius distance (Frob. dist.) to the ground-truth scDNA correlation is computed. **c**, Left: the empirical correlation matrix of scDNA data across clusters as ground-truth clonal dependencies. The color bar shows the reported grouping of clusters into ground-truth clone structure[40]. Right: the learned correlation matrix from Echidna. The color bar shows clones identified by Echidna through hierarchical clustering and tree cutting. The dendrogram represents the learned tree grouping cell clusters. **d**, UMAPs of the scRNA-seq data colored by clones identified by Echidna, InferCNV and Numbat, respectively. **e**, Benchmarking clone reconstruction across all models. The metrics used are purity, AMI and ARI. Box plots show the distribution of each metric across $n = 100$ bootstrap resamples of the cell population. Center line, median; box limits, 25th–75th percentile; whiskers, 1.5 × IQR. Each method was compared to Echidna using a two-sided independent-samples t-test (d.f. 198; $n_1 = n_2 = 100$ bootstrap samples per group). P values for each pairwise comparison are reported above each box. 'XClone-allele' represents running XClone only with the BAF module that observes allelic information, while 'XClone' means running the full XClone model.

https://doi.org/10.1038/s41592-026-03279-4

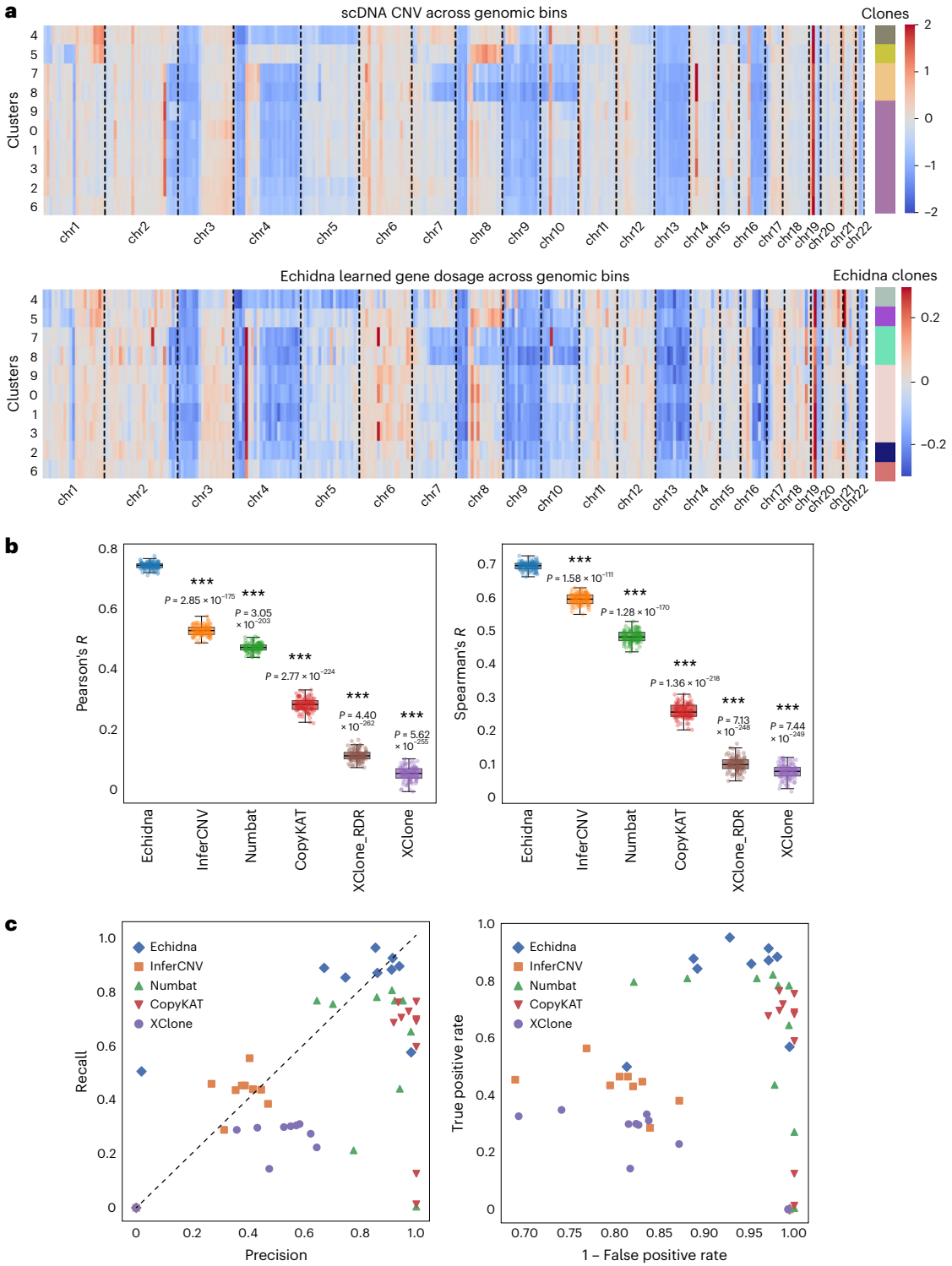


Fig. 3 | Echidna accurately estimates cluster-level CNA states and gene dosage. a, Cluster-level gene dosage profile from an endometrial cancer sample (tumor 2 from ref. 40) showing ground-truth scDNA collected via hybrid snDNA-seq (top) and inferred from hybrid scRNA-seq and matched WGS by Echidna (bottom). **b**, Benchmarking gene dosage estimation across models. Pearson and Spearman correlations are computed against ground-truth binned scDNA data. Box plots show the distribution of each metric across $n = 100$ bootstrap resamples of the cell population. Center line, median; box limits, 25th–75th percentile; whiskers, 1.5 × IQR. Each method was compared to Echidna using a two-sided independent-samples t-test (d.f. 198; $n_1 = n_2 = 100$ bootstrap samples per group). P values for each pairwise comparison are reported above each box. **c**, CNA state prediction performance by different methods. Each dot represents a cluster (phenotypic state).

on the complete dataset, after which semi-synthetic observations were generated from its inferred generative process to establish ground-truth parameters. The model was then refit under two perturbations: (1) systematic variation of key hyperparameters and (2) progressive subsampling of the synthetic data. Robustness was quantified by computing the Pearson correlation or Frobenius distance between the refitted posterior parameters and their ground-truth counterparts. We observe that Echidna is robust across various

choices of hyperparameters and can handle low sample size regimes (Extended Data Fig. 3a–c).

Echidna achieves robust estimation of gene dosage and accurate prediction of CNV states

In addition to evaluating performance in clonal structure, we benchmarked Echidna's ability to estimate CNA profiles at the level of phenotypic clusters. Echidna's gene dosage estimates revealed distinct clone-specific CNA events that match ground-truth scDNA profiles (Fig. 3a). Specifically, we note amplifications in chromosome 1 in clusters 4 and 5; amplifications in chromosome 4 in cluster 5; and deletions in chromosomes 7, 8 and 9 in clusters 7 and 8 (which were previously identified to be from the same clone; Fig. 2c).

To evaluate the quality of gene dosage estimates, we compared predicted cluster-level gene dosage estimates from each method to the scDNA ground truth. We utilize similar preprocessing of the scDNA data as ref. 40, binning the data into 300 genomic bins to improve the signal. We obtained cluster-level CNA estimations per gene from all methods and grouped them to match these genomic bins[40]. For Echidna, we directly used the posterior mean of our gene dosage random variable. Gene dosage estimates for other methods were obtained as described in (Methods). Reconstruction quality was assessed using Pearson correlation coefficient (Pearson's *R*) and Spearman correlation coefficient (Spearman's *R*) between estimated gene dosage profiles and the scDNA ground truth. Echidna outperforms all methods in gene dosage profile estimation and was able to detect more accurate clone-specific patterns (Fig. 3a,b and Extended Data Fig. 4a–c).

To evaluate the copy number variation (CNV) state prediction performance within each cell cluster, we determined ground-truth state labels by applying Gaussian HMMs to the scDNA gene dosage profiles for each cluster across genomic bins. Each bin was categorized as amplification, deletion or neutral based on the emitted HMM states. We obtain Echidna's predicted CNV states by the same post-processing steps described in the overview. We assign states to the cluster-level gene dosage profiles estimated by InferCNV and CopyKAT by applying thresholds based on the standard deviations within each cluster, while Numbat and XClone CNV states were acquired directly from the MAP estimates. We calculated precision, recall, true positive rate and false positive rate to systematically assess CNV state prediction performance for each cell cluster. We observe superior performance by Echidna in predicting CNV states in most clusters compared to all other methods (Fig. 3c, left; Methods). Echidna also achieves higher true positive rates while maintaining low false positive rates (Fig. 3c, right).

Altogether, these results highlight the advantage of integrating scRNA with WGS data, where the former ensures accountability for phenotypic variability and the latter serves as a constraint on cluster-specific CNA inference. This unified framework enables accurate estimation of gene dosage and its downstream transcriptional effects.

Echidna accurately detects clones characterized by focal amplifications

We next applied Echidna to matched snRNA-seq and snDNA-seq data with paired WGS from a glioblastoma specimen characterized by two distinct clones harboring amplifications in key genes, specifically *EGFR*, *PDGFRA* and *MDM2* (ref. 32). This dataset, collected using the DEFND-seq platform, provided an ideal validation scenario, as small-scale focal amplifications are often challenging to detect due to the inherent smoothing and binning strategies used by most CNV inference tools[32]. Furthermore, this sample featured extrachromosomal amplifications that greatly skewed the range of possible CNVs, with genes such as *EGFR* approaching close to 100 copies, as determined by bulk genomic sequencing (Extended Data Fig. 5a). Ground-truth CNV states derived from snDNA-seq served as the benchmark for evaluating Echidna's accuracy (Extended Data Fig. 5b).

We first demonstrated that Echidna successfully captured clonal structure defined by *EGFR* and *PDGFRA* as described in ref. 32 (Fig. 4a). These clones align with expected increased mRNA expression of *EGFR* and *PDGFRA*, respectively (Fig. 4b). Echidna outperforms other methods in CNV reconstruction benchmarks against binned snDNA-seq (Fig. 4c). Echidna additionally captures key focal amplifications (Fig. 4d), with global amplifications in *MDM2* and expected clone-specific amplifications in *EGFR* and *PDGFRA* (Fig. 4e). Differential expression analysis further confirmed that clone C2 exhibited the expected downregulation of *EGFR* and upregulation of *PDGFRA* (Fig. 4f).

Echidna captures phenotypic diversity while preserving genetic similarity in metastatic niches

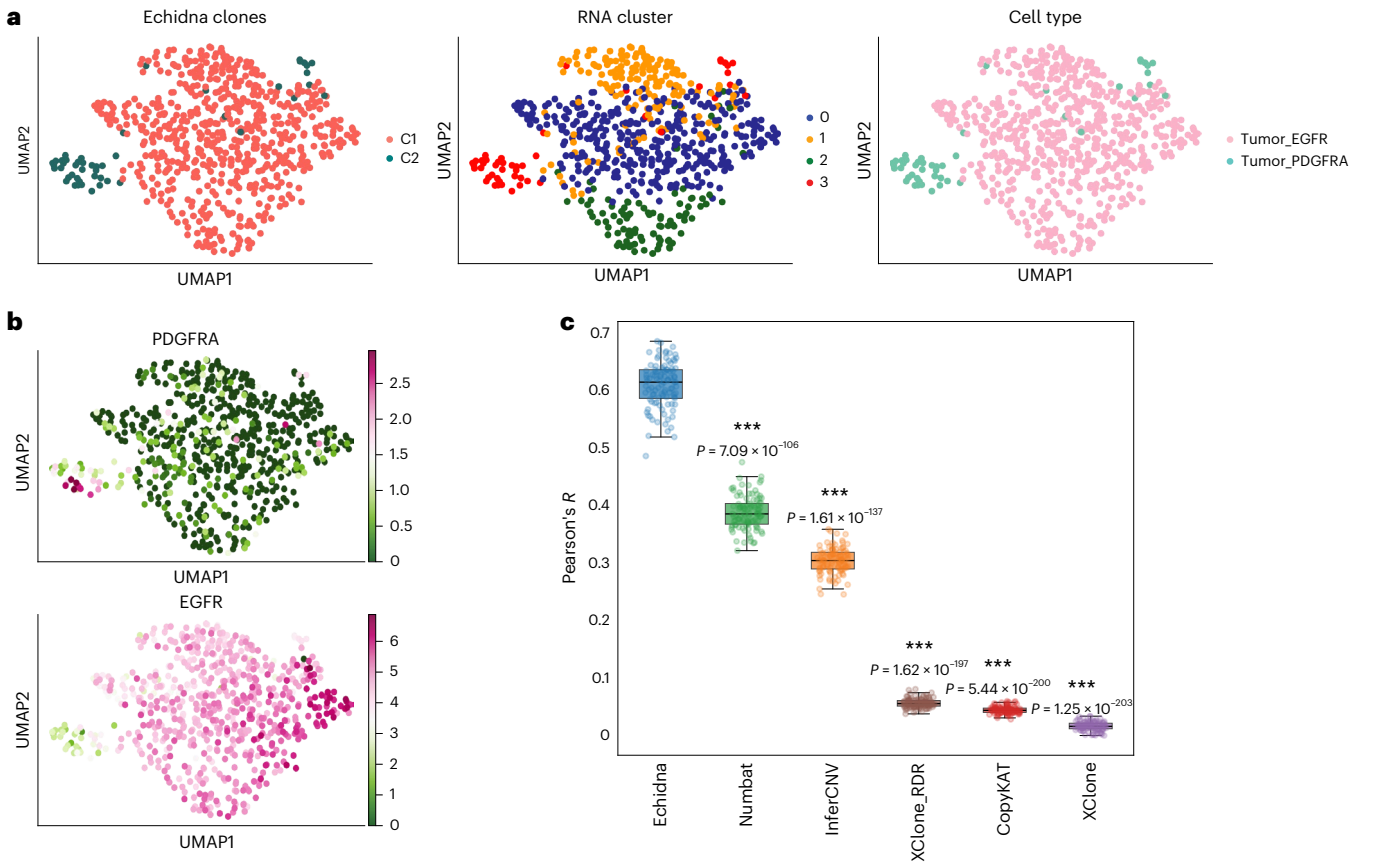
As Echidna relies upon the integration of WGS with transcriptomic data to accurately infer CNAs, we aimed to investigate whether pronounced differences in gene expression between clones may lead to spurious calls in CNAs. We evaluated Echidna on a cohort of pancreatic adenocarcinoma (PDAC) primary tumors with eventual recurrent liver or lung metastases[44]. Critically, although primary tumors with eventual liver versus lung metastatic recurrences showed distinct upregulation in transcriptional programs corresponding to their future metastatic niche, there was no detectable genomic signature present[44]. Echidna recapitulates these findings, demonstrating that of the 632 genes in the liver signature and the 999 genes of the lung signature, only 11 and 3 genes have significantly different CNAs as inferred by Echidna, respectively (Extended Data Fig. 6; Methods). Thus, Echidna does not detect spurious differences in CNAs between key conditions, even when such differences exist at the transcriptomic level.

Echidna characterizes clonal evolution during tumor metastasis

To understand the extent to which phenotypic variability over time is enabled by gene dosage, Echidna integrates scRNA-seq and WGS data from different time points in the same system, linking clonal evolution to temporal gene expression dynamics. We first demonstrate this integrative capability in the setting of tumor metastasis, with the goal of identifying primary tumor clones and enabling CNAs that probably gave rise to metastatic phenotypes. We applied this framework to a primary non-small cell lung cancer (NSCLC) tumor and a matched brain metastasis[45]. Cells from both samples were first grouped into phenotypic clusters (Fig. 5a). Analysis of inferred CNAs and cluster correlations (Fig. 5b) identified a metastasis-associated clone (C1) characterized by amplifications across chromosomes 2, 4, 5, 7, 9, 16 and 20. Primary clones C2 and C3 were most similar to C1, while C4 grouped with a second metastasis-associated clone (C5). These results suggest two potentially distinct mechanisms of metastasis, with the dominant

Fig. 4 | Echidna detects focal amplifications in glioblastoma. a, UMAP projection of snRNA-seq on glioblastoma tumor cells collected by ref. 32, colored by Echidna-inferred clones (left), phenotypic clusters (middle) and ground-truth clones defined from matched snDNA-seq (right). **b**, UMAP representation of glioblastoma tumor cells colored by expression of key receptors *EGFR* and *PDGFRA*. **c**, Benchmarking gene dosage estimation across models. Pearson correlation is computed against ground-truth binned scDNA data from ref. 32. Box plots show the distribution of each metric across $n = 100$ bootstrap resamples of the cell population. Center line, median; box limits, 25th–75th percentile; whiskers, 1.5 × IQR. Each method was compared to Echidna using a two-sided independent-samples *t*-test (d.f. 198; $n_1 = n_2 = 100$ bootstrap samples per group). *P* values for each pairwise comparison are reported above each box. **d**, Echidna-inferred CNA for each cluster of glioblastoma tumor cells. Known focal amplifications are labeled in red and focal deletions in blue. **e**, Inferred CNAs of known focally amplified/deleted genes. **f**, Differentially downregulated/upregulated genes in C2, colored by *z*-scored expression across clusters.

https://doi.org/10.1038/s41592-026-03279-4



mechanism encompassing C1–C3 revealing large-scale CNA events at the heart of the clonal transformation.

To identify genomic drivers of metastasis, we first computed DEGs between metastasized and primary tumor cells. Then, we examined the density of upregulated genes along the genome to find those concentrated in metastatic-specific regions of amplification or deletion. We label these regions as 'hotspots' and interpret them as candidate loci of functional alteration. Gene set enrichment, computed as a Fisher's exact test (Methods), on hotspot DEGs regions reveals pathways related to epithelial–mesenchymal transition as well as inflammation and hypoxia (Fig. 5c).

We further identified genes that are both upregulated in metastatic clones and exhibit significant CNA differences relative to primary clones (Fig. 5d,e; Methods). Among these genes are *APP*, implicated in brain metastasis in melanoma[46]; *TGIF1* and *TRIM27*, which have known roles in the metastasis of NSCLC[47,48]; *CDK12*, *GPNMB*, *HNRNPA2B1*, *RALA* and *TSG101*, with demonstrated metastatic function in other cancers[49–53]; and *SRSF1*, with roles in tumor invasiveness[54]. These results suggest that metastasis of C1 may have in part been driven by advantageous CNA gains in key genes, and highlight the importance of integration across samples in the examination of genomic drivers of tumor evolution.

Echidna tracks clonal evolution during immune checkpoint blockade therapy using longitudinal data

Understanding the role of genetically driven resistance in a therapeutic context often requires establishing a longitudinal context so that evolving phenotypes may be examined over time. Therefore, we applied Echidna to longitudinal melanoma data from a patient treated with anti-PD-1 therapy[22]. Tumor cells from these data exhibit pronounced heterogeneity between the pre-therapy and on-therapy specimens (with little overlap in clusters)[22] (Fig. 6a). This heterogeneity is not due to technical (batch) effects, as shared immune cell states were detected between the time points[22]. We first confirmed that Echidna accurately reconstructs both scRNA-seq and WGS data across the two time points (Extended Data Fig. 7a,b), showing that the multi-time point model effectively fits the data.

Analysis of inferred CNAs reveals diverse amplification and deletion patterns that define clonal structure (Extended Data Figs. 1b and 7c,d) while reconstructing gene expression data (Extended Data Fig. 7e). Echidna recovers a clone shared between the two time points (clone 3), previously reported in this patient, which expands under treatment[22] (Fig. 6a). This clone, marked by amplifications in chromosomes 1 and 6, is rare pre-treatment (cluster 9) and expands post-treatment (cluster 5; Fig. 6a). Clone 2, sharing similar amplifications, is only present on treatment, suggesting potentially shared resistance mechanisms. By contrast, clones 1 and 4 contract with treatment. Compared to existing methods, only inferCNV and XClone were able to correctly identify the expanding clone, while Numbat and CopyKat did not (Extended Data Figs. 8 and 9). Echidna's overall performance in recovering clonal architecture from time series data was further demonstrated to be robust to small cluster sizes on semi-synthetic data (Supplementary Note B and Supplementary Fig. 3a–c).

Echidna quantifies the effect of gene dosage on temporal gene expression dynamics and phenotypic plasticity

We next utilize the GDX metric, which quantifies the proportion of transcriptional rate explained by CNAs, to gain deeper insight into mechanisms driving the clonal expansion phenotype. To identify drivers of treatment response, we first defined DEGs between cells from expanding and contracting clones. By examining the genomic density of DEGs, we identified hotspots of differential gene activity concentrated in amplified regions, associated with expanding and contracting clones (Fig. 6b and Extended Data Fig. 7f, respectively). The top DEGs within these hotspots showed significantly higher GDX ($P < 1 \times 10^{-3}$ for all clusters) compared to non-hotspot DEGs, verifying that their increased expression is driven by underlying amplifications.

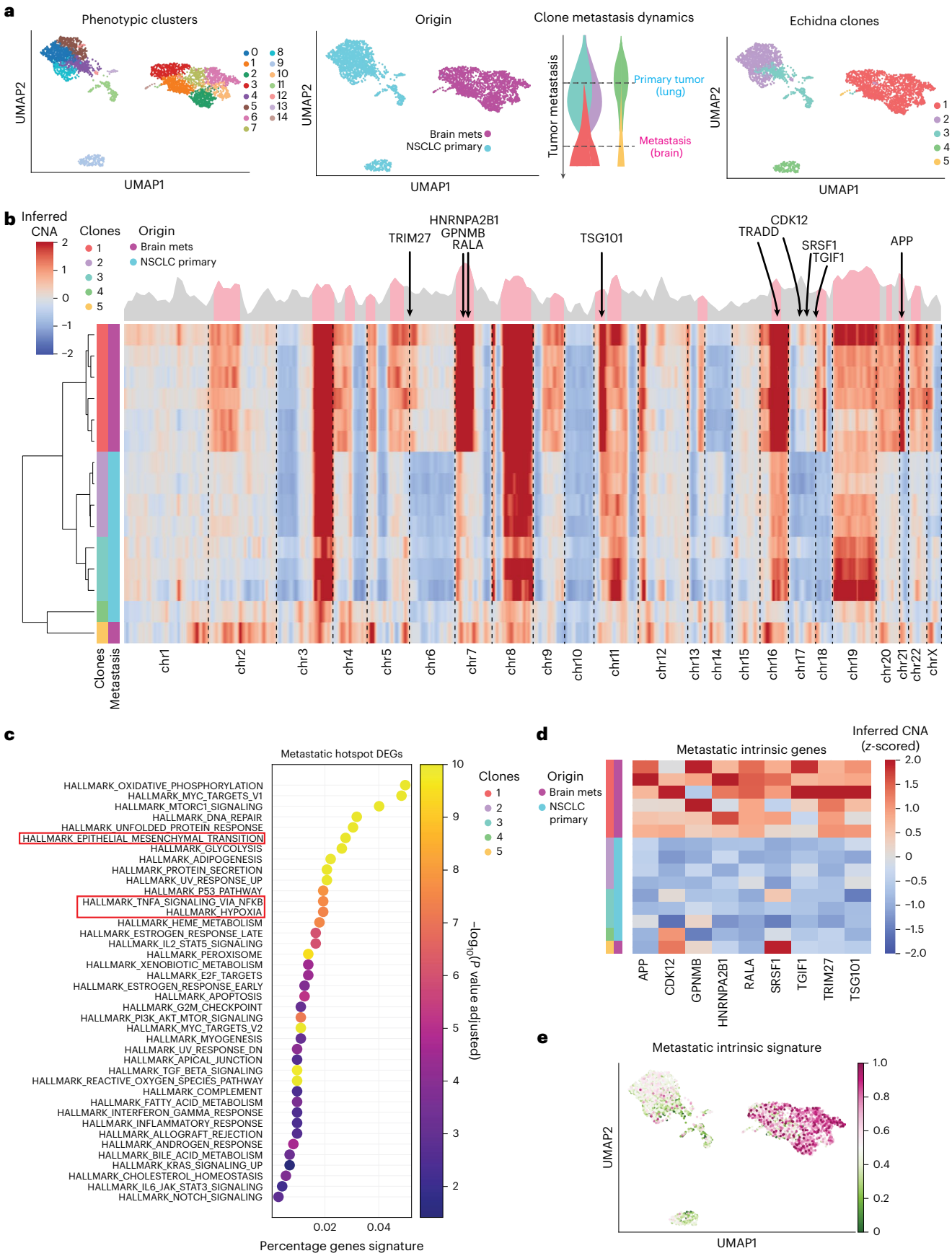
In the contracting clones, we identified four separate hotspot regions (Extended Data Fig. 7f), containing zinc finger proteins (for example, *ZFP64*, *ZNF208*, *ZNF91*, *ZNF790-AS1*, *ZNF506*, *ZNF875*, *ZNF529* and *ZNF146*) and long intergenic noncoding RNAs (for example, *LINC02438*, *LINC02267*, *LINC01531*, *LINC00906* and *LINC01524*), several of which have known roles in melanoma progression and immunotherapy response[55–59].(Supplementary Note C). In expanding clones, hotspot H1 (Fig. 6b,c) contains multiple *S100A* family genes, which is consistent with this family of genes' previously established role in early diagnostics of melanoma[60,61]. A second hotspot, H2, was highly enriched in MHC-II genes, reflecting recent discussions on the complex role of MHC-II in melanoma response to immunotherapy[62]. Our results demonstrate that these genes are not only found to be upregulated in resistant clones, but they are also linked with underlying genomic amplifications.

We next examined the role of specific immune checkpoint resistance (ICR) genes as defined in refs. 63,64 (Fig. 6b and Extended Data Fig. 7f). We demonstrated that our GDX score is able to distinguish between subsets of ICR genes more robustly than gene expression. Specifically, we defined two sets of genes for statistical testing: ICR hotspot genes upregulated in expanding clones and non-hotspot ICR genes upregulated in expanding clones. These two gene sets were then examined in both expanding and contracting clusters. We show that the GDX is significantly different not only between expanding and contracting clusters, but also between hotspot and non-hotspot genes within categories. Conversely, gene expression only captures the difference between expanding and contracting clones (Fig. 6d). At the pathway level, GDX again demonstrates increased granularity, revealing additional differences in mesenchymal, metastatic and immune resistance signatures that are not captured by expression alone[65] (Supplementary Fig. 1a,b). Notably, GDX distinguishes between expanding clones, highlighting that only the clone consisting of clusters 5 and 9 is intrinsically driven by mesenchymal signatures.

Finally, we interpret our GDX metric in the spatial context using previously published spatial transcriptomic (SlideSeq) data[22]. We first highlight that clones identified by Echidna appear as spatially distinct regions in the tumor (Fig. 6e). Specifically, clone 4, which expresses the ICR signature[22], is spatially separated from clone 2, suggesting that clonal expansion occurred in a partitioned fashion. Moran's *I* spatial autocorrelation analysis[66] identified genes with distinct spatial localization, including *S100A4* (GDX of 0.59), which is associated with tumor metastasis through epithelial–mesenchymal transition[67], and *HLA-DRA* (GDX of 0.4), an MHC class II family gene. The higher GDX scores suggest that genetic events for these genes helped drive spatial variability. By contrast, genes such as *CD74* (GDX of 0.11) have a high spatial autocorrelation but low GDX, suggesting spatial variation probably driven by extrinsic factors. These results further suggest that

Fig. 5 | Echidna integrates across the metastatic axis, finding CNAs relevant to tumor metastasis. a, UMAP computed from snRNA-seq on primary NSCLC tumor cells combined with brain metastases cells from ref. 45, colored by phenotypic clusters (left), primary or metastatic origin (middle) and Echidna-inferred clones (right). **b**, Inferred CNAs for each phenotypic cluster, grouped according to hierarchical clustering based on correlation. Density above indicates the location of DEGs characterizing metastasized clones, with red regions indicating hotspots. Key genes from **d** are labeled in black. **c**, Gene set enrichment of GSEA Hallmark genes using Fisher's exact test, computed on hotspot genes from **b**. Color indicates the −log10(BH-adjusted *P* value). **d**, *z*-scored inferred CNA values for DEGs in metastasized clones with statistically significant change in CNA between metastasized and primary clusters, limited to genes relevant to metastasis/invasion. **e**, UMAP computed from **a** colored by the mean expression of relevant intrinsic genes in **d**.

https://doi.org/10.1038/s41592-026-03279-4



https://doi.org/10.1038/s41592-026-03279-4

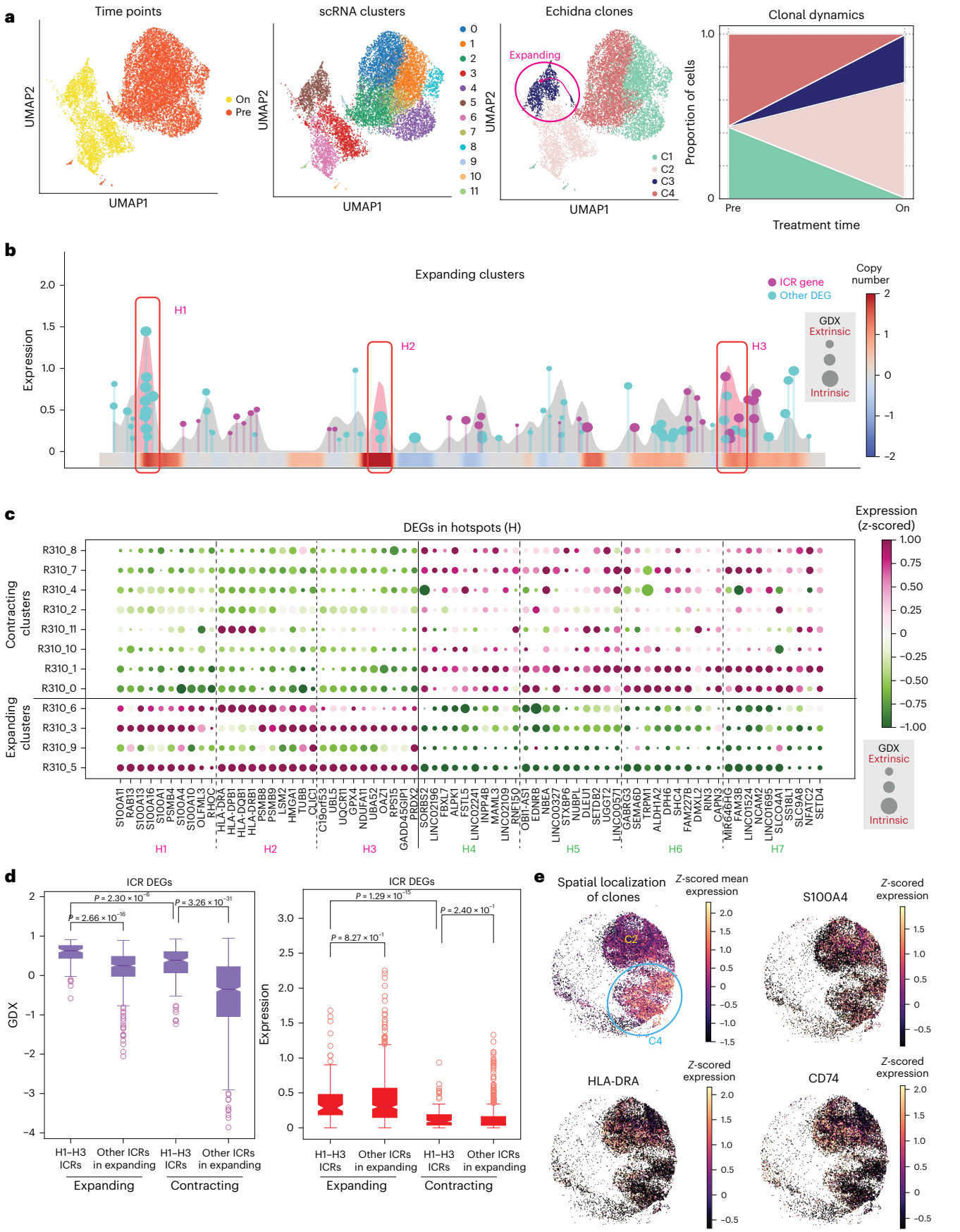


Fig. 6 | GDX reveals mechanisms underlying clone-specific response to treatment. a, UMAP projection of scRNA-seq of tumor cells combined across samples, colored by treatment time point: pre- and on-immunotherapy (first) and phenotypic (Leiden) clusters (second), clonal structure revealed through correlation (third). The proportion of cells belonging to each Echidna clone at each time point (fourth). **b**, Top DEGs in clusters expanding with treatment, visualized as lollipops with the *x* axis corresponding to genomic position. Density plot reflects the distribution of DEGs by genomic position. Height of each lollipop corresponds to the expression of each gene, while dot size reflects the absolute value of the GDX, such that larger dots denote stronger dependency of expression on gene dosage. Mean gene dosage is plotted underneath as a heat map. ICR genes are highlighted in pink. **c**, Dot plot of top DEGs in the hotspots highlighted in **b** and Extended Data Fig. 7f for each cluster. Dot size corresponds to the absolute value of GDX, while color represents *z*-scored expression. **d**, GDX values (left) and expression (right) for ICR genes upregulated in expanding clones (resistant to treatment, 23 genes) that are located in genomic hotspots H1–H3, and those not in any hotspots (117 genes); in expanding (3,821 cells) and contracting (9,577 cells) clones. *P* values were computed using the two-tailed Mann–Whitney *U* test. Center line represents median GDX/expression; lower box limit represents quartile 1 and upper box limit represents quartile 3; whiskers demonstrate 1.5× the IQR; and points are any points beyond the whiskers. **e**, Spatial transcriptomic (Slideseq) data for the on-treatment sample. Spots are colored by the expression of clone 4 signature (also circled) with the location of clone 2 labeled, and expression of key genes with high spatial variation.

Echidna may unveil underlying dynamics governing the transformation and localization of complex cell lineages.

Discussion

Phenotypic plasticity has emerged as an important concept in tumorigenesis, tumor progression and resistance to targeted and immune-based treatments. While advances in scRNA-seq have revealed extensive transcriptional heterogeneity, a major challenge in the field has been the lack of methods that accurately estimate and account for gene dosage across multiple time points and conditions. Understanding whether transcriptional changes are merely a consequence of gene dosage differences between cancer clones within a sample or across patient cohorts, and separating this effect from putative plasticity induced through interactions with other cells of the tumor microevironment or nongenomically encoded events, remains a major unmet need in the analyses of tumor data.

While concurrent measurement of scDNA- and RNA-seq could address some of these challenges, coupling these methods remains a technically challenging task and is difficult to scale. By contrast, obtaining bulk WGS and scRNA-seq from the same specimen is already feasible and scalable. We therefore developed Echidna, a computational framework that integrates these modalities to improve estimation of gene dosage and quantify its contribution to transcriptional variability at scale, and over time, within and across patient samples.

The conceptual core of Echidna lies in its hierarchical modeling of the relationship between clonal identity and transcriptional programs, enabling a fine-grained dissection of cellular adaptation and clonal dynamics when data from multiple time points are available. In the context of cancer, where clonal diversity often drives therapy resistance and relapse, Echidna provides a crucial tool for studying these processes at high resolution and in a technically and computationally scalable format. Given the emerging recognition that large chromosomal aberrations rather than specific point mutations may determine drug resistance in different disease and treatment contexts[68], Echidna provides a framework to begin unraveling underlying mechanisms while accounting for different sources of plasticity.

In its current form, Echidna is applicable to cancer types where CNAs are a defining hallmark. However, Echidna does not currently account for transcriptional changes as a result of epigenetic regulation. Understanding the complex interplay between chromatin accessibility, CNA and phenotypic plasticity can be conceivably evolved in the future through integration of additional data analytes. Similarly, SNVs may also be detected from WGS data and built into the prior for cluster similarities to improve model performance. In addition, the GDX metric introduced here could be extended to more rigorously distinguish transcriptional programs driven by genomic alterations from those arising through other mechanisms. Together, these advances will further enhance our ability to identify genomic drivers of phenotypic plasticity.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03279-4.

References

1. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
2. Padovan-Merhar, O. et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
3. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* **22**, 3–18 (2021).
4. Kakiuchi, N. & Ogawa, S. Clonal expansion in non-cancer tissues. *Nat. Rev. Cancer* **21**, 239–256 (2021).
5. Sade-Feldman, M. et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).
6. Lauss, M. et al. Molecular patterns of resistance to immune checkpoint blockade in melanoma. *Nat. Commun.* **15**, 3075 (2024).
7. Yeh, C. Y. et al. Mapping spatial organization and genetic cell-state regulators to target immune evasion in ovarian cancer. *Nat. Immunol.* **25**, 1943–1958 (2024).
8. Martínez-Ruiz, C. et al. Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature* **616**, 543–552 (2023).
9. Angerilli, V. et al. Intratumor morphologic and transcriptomic heterogeneity in V600EBRAF-mutated metastatic colorectal adenocarcinomas. *ESMO Open* **6**, 100211 (2021).
10. Williams, M. J. et al. Luminal breast epithelial cells of BRCA1 or BRCA2 mutation carriers and noncarriers harbor common breast cancer copy number alterations. *Nat. Genet.* **56**, 2753–2762 (2024).
11. Fischer, U., Kim, E., Keller, A. & Meese, E. Specific amplifications and copy number decreases during human neural stem cells differentiation towards astrocytes, neurons and oligodendrocytes. *Oncotarget* **8**, 25872 (2017).
12. Liang, Z. et al. High chromosome instability identified by low-pass whole-genome sequencing assay is associated with TP53 copy loss and worse prognosis in BRCA1 germline mutation breast cancer. *Breast Cancer* **29**, 103–113 (2022).
13. Iacobuzio-Donahue, C. A., Velculescu, V. E., Wolfgang, C. L. & Hruban, R. H. Genetic basis of pancreas cancer development and progression: insights from whole-exome and whole-genome sequencing. *Clin. Cancer Res.* **18**, 4257–4265 (2012).
14. Nowak, M. A. et al. The role of chromosomal instability in tumor initiation. *Proc. Natl Acad. Sci. USA* **99**, 16226–16231 (2002).
15. Bristy, N. A., Fu, X. & Schwartz, R. Sc-TUSV-Ext: single-cell clonal lineage inference from single nucleotide variants, copy number alterations, and structural variants. *J. Comput. Biol.* **32**, 723–737 (2025).

16. Fu, X., Lei, H., Tao, Y. & Schwartz, R. Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations. *Bioinformatics* **38**, i125–i133 (2022).
17. Jun, S. H. et al. Reconstructing clonal tree for phylo-phenotypic characterization of cancer using single-cell transcriptomics. *Nat. Commun.* **14**, 982 (2023).
18. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
19. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
20. Bärthel, S., Falcomatà, C., Rad, R., Theis, F. J. & Saur, D. Single-cell profiling to explore pancreatic cancer heterogeneity, plasticity and response to therapy. *Nat. Cancer* **4**, 454–467 (2023).
21. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
22. Wang, Y. et al. Multimodal single-cell and whole-genome sequencing of small, frozen clinical specimens. *Nat. Genet.* **55**, 19–25 (2023).
23. Maurer, A. et al. Coordinated immune cell networks in the bone marrow microenvironment define the graft versus leukemia response with adoptive cellular therapy. *Sci. Immunol.* **10**, eadr0782 (2025).
24. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
25. Dong, D. et al. An RNA-informed dosage sensitivity map reflects the intrinsic functional nature of genes. *Am. J. Hum. Genet.* **110**, 1509–1521 (2023).
26. Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
27. Akavia, U. D. et al. An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
28. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221 (2019).
29. Shi, H. et al. Allele-specific transcriptional effects of subclonal copy number alterations enable genotype-phenotype mapping in cancer cells. *Nat. Commun.* **15**, 2482 (2024).
30. Weiner, A. C. et al. Inferring replication timing and proliferation dynamics from single-cell DNA sequencing data. *Nat. Commun.* **15**, 8512 (2024).
31. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).
32. Olsen, T. R. et al. Scalable co-sequencing of RNA and DNA from individual nuclei. *Nat. Methods* **22**, 477–487 (2025).
33. Bhattacharya, A. et al. Transcriptional effects of copy number alterations in a large set of human cancers. *Nat. Commun.* **11**, 715 (2020).
34. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
35. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
36. Gao, T. et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat. Biotechnol.* **41**, 417–426 (2023).
37. Nazaret, A. et al. Joint representation and visualization of derailed cell states with Decipher. *Genome Biol.* **26**, 219 (2025).
38. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
39. Lewandowski, D., Kurowicka, D. & Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001 (2009).
40. Li, S. et al. Hybrid BAG-seq: DNA and RNA from the same single nucleus reveals interactions between genomic and transcriptomic landscapes in human tumor samples. *Genome Biol.* **26**, 314 (2025).
41. Tickle, T. et al. inferCNV of the Trinity CTAT Project. *GitHub* https://github.com/broadinstitute/infercnv (2019).
42. Gao, R. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).
43. Huang, R. et al. Robust analysis of allele-specific copy number alterations from scRNA-seq data with XClone. *Nat. Commun.* **15**, 6684 (2024).
44. Chalabi Hajkarim, M. et al. Cellular states associated with metastatic organotropism and survival in patients with pancreatic ductal adenocarcinoma. *Nat. Genet.* **57**, 2728–2742 (2025).
45. Tagore, S. et al. Single-cell and spatial genomic landscape of non-small cell lung cancer brain metastases. *Nat. Med.* **31**, 1351–1363 (2025).
46. Kleffman, K. et al. Melanoma-secreted amyloid beta suppresses neuroinflammation and promotes brain metastasis. *Cancer Discov.* **12**, 1314–1335 (2022).
47. Xiang, G., Yi, Y., Weiwei, H. & Weiming, W. TGIF1 promoted the growth and migration of cancer cells in nonsmall cell lung cancer. *Tumor Biol.* **36**, 9303–9310 (2015).
48. Liu, S. et al. TRIM27 acts as an oncogene and regulates cell proliferation and metastasis in non-small cell lung cancer through SIX3-β-catenin signaling. *Aging* **12**, 25564 (2020).
49. Lu, K. Q. et al. CDK12 is a potential biomarker for diagnosis, prognosis and immunomodulation in pan-cancer. *Sci. Rep.* **14**, 6574 (2024).
50. Maric, G., Rose, A. A. N., Annis, M. G. & Siegel, P. M. Glycoprotein non-metastatic b (GPNMB): a metastatic mediator and emerging therapeutic target in cancer. *Onco Targets Ther.* **6**, 839–852 (2013).
51. Zhu, F., Yang, T., Yao, M., Shen, T. & Fang, C. HNRNPA2B1, as a m6A reader, promotes tumorigenesis and metastasis of oral squamous cell carcinoma. *Front. Oncol.* **11**, 716921 (2021).
52. Thies, K. A. et al. The small G-protein RalA promotes progression and metastasis of triple-negative breast cancer. *Breast Cancer Res.* **23**, 65 (2021).
53. Yang, Y., Wang, X. L., Yue, Y. X., Chen, G. & Xia, H. F. TSG101 overexpression enhances metastasis in oral squamous cell carcinoma through cell cycle regulation. *Cell. Signal.* **125**, 111519 (2025).
54. Li, S. et al. High expression of SRSF1 facilitates osteosarcoma progression and unveils its potential mechanisms. *BMC Cancer* **24**, 580 (2024).
55. Melixetian, M. et al. Regulation of lncRNAs in melanoma and their functional roles in the metastatic process. *Cells* **11**, 577 (2022).
56. Czerwinska, P., Jaworska, A. M., Wlodarczyk, N. A. & Mackiewicz, A. A. Melanoma stem cell-like phenotype and significant suppression of immune response within a tumor are regulated by TRIM28 protein. *Cancers* **12**, 2998 (2020).
57. Machnik, M. et al. The expression signature of cancer-associated KRAB-ZNF factors identified in TCGA pan-cancer transcriptomic data. *Mol. Oncol.* **13**, 701–724 (2019).
58. Siddiqui, J. & Miles, W.O. RNA editing signatures identify melanoma patients who respond to pembrolizumab or nivolumab treatment. *Translational Oncol.* **14**, 101197 (2021).

59. Wozniak, M. & Czyz, M. The functional role of long non-coding RNAs in melanoma. *Cancers* **13**, 4848 (2021).
60. Gaynor, R. et al. S100 protein: a marker for human malignant melanomas? *Lancet* **317**, 869–871 (1981).
61. Xiong, T. F., Pan, F. Q. & Li, D. Expression and clinical significance of S100 family genes in patients with melanoma. *Melanoma Res.* **29**, 23–29 (2019).
62. Johnson, D. B. et al. Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nat. Commun.* **7**, 10582 (2016).
63. Jerby-Arnon, L. et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* **175**, 984–997.e24 (2018).
64. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
65. Wouters, J. et al. Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. *Nat. Cell Biol.* **22**, 986–998 (2020).
66. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
67. Fei, F., Qu, J., Zhang, M., Li, Y. & Zhang, S. S100A4 in cancer progression and metastasis: a systematic review. *Oncotarget* **8**, 73219 (2017).
68. Kou, F., Wu, L., Ren, X. & Yang, L. Chromosome abnormalities: new insights into their clinical significance in cancer. *Mol. Ther. Oncolytics* **17**, 562–570 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

[1]Department of Biomedical Engineering, Columbia University, New York, NY, USA. [2]Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA. [3]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA. [4]Department of Molecular Pharmacology and Therapeutics, Columbia University Medical Center, New York, NY, USA. [5]Department of Applied Mathematics, Columbia University, New York, NY, USA. [6]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA. [7]Department of Computer Science, Columbia University, New York, NY, USA. [8]Immunology program, Stanford University, Stanford, CA, USA. [9]Department of Medicine, Division of Hematology and Oncology, Columbia University Irving Medical Center, New York, NY, USA. [10]Columbia Center for Translational Immunology, Columbia University Irving Medical Center, New York, NY, USA. [11]Biohub, New York, NY, USA. [12]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [13]Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, USA. [14]Program for Mathematical Genomics, Columbia University, New York, NY, USA. [15]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA. [16]These authors contributed equally: Joy Linyue Fan, Mingxuan Zhang. ✉e-mail: bi2175@cumc.columbia.edu; ea2690@columbia.edu

Methods

Model specification

Notations. Echidna is a Bayesian hierarchical model. We utilize variational inference to infer posterior distributions of latent random variables that generate observed variables through a designed generative process. Assume we have T time points, consider the WGS copy number vector $W \in \mathbb{Z}_+^{G\times T}$ across G genes and the scRNA count matrix $X \in \mathbb{Z}_+^{T\times N\times G}$ for N cells, where $\mathbb{Z}_+$ denotes a set of non-negative integers. We denote the deconvolved gene dosage (copy number) in the K cell clusters as $\eta \in \mathbb{R}^{G\times K}$ and the time-specific latent cell transcriptional rate as $c \in \mathbb{R}_+^{T\times G\times K}$, where $\mathbb{R}$ is the set of real numbers and $\mathbb{R}_+$ is the set of positive real numbers. We also denote the proportion of observed cell clusters as $\pi \in \mathbb{R}^{T\times K}$ and the cluster membership matrix as $z \in [0,1]^{T\times N\times K}$. The library size (sum of counts per cell in scRNA-seq) is denoted as $l \in \mathbb{Z}_+^{T\times N}$. We denote K dimensional multivariate normal distributions as $\mathcal{N}^K$.

Generative process of Echidna. We build correlation structures between clusters of cells (within or across time points) informed by both scRNA-seq and WGS. Clusters with highly correlated ηs are interpreted as clones (which can be static and over time). Therefore, we treat the covariance matrix of η as a random variable. The lower Cholesky factors L of the covariance matrix Σ is given by

$$L \sim \mathrm{LKJCholesky}(K, \beta), \tag{1}$$

where β is the concentration parameter. For each cluster k, we sample cluster variances σ_η from inverse Gamma distributions as

$$\boldsymbol{\sigma}_\eta[k] \sim \Gamma^{-1}(1,1), \quad \boldsymbol{\sigma}_\eta \in \mathbb{R}^K \tag{2}$$

We combine the Cholesky factors and the diagonal variances to parametrize the full covariance of η as

$$\Sigma = LL^{\mathrm{T}} \tag{3}$$

and updated the diagonal as

$$\Sigma \leftarrow \Sigma - I + \mathrm{diag}(\boldsymbol{\sigma}_\eta) \tag{4}$$

where we obtain the full covariance with unit variance in equation (3) and scale the diagonal (diag) elements with the sampled variance σ_η in equation (4). We sample the mean of η from a K dimensional multivariate normal centered around neutral copy number (2) such that

$$\boldsymbol{\mu}_\eta \sim \mathcal{N}^K(\mathbf{2}, I), \quad \boldsymbol{\mu}_\eta \in \mathbb{R}^K \tag{5}$$

We model gene dosage as deconvolved copy number profiles across K clusters for all genes as G different K dimensional sampling distributions. The gene dosage of gene i is given by

$$\boldsymbol{\eta}_i \sim \mathcal{N}^K(\boldsymbol{\mu}_\eta[i], \Sigma) \tag{6}$$

$$\boldsymbol{\eta}_i \leftarrow \mathrm{softplus}(\boldsymbol{\eta}_i), \quad \boldsymbol{\eta}_i \in \mathbb{R}^K \tag{7}$$

where in equation (7) we apply a softplus activation to the sampled gene dosage to keep it positive. Assuming gene dosage in a specific clone does not change dramatically over time, the observed copy number W_i for gene i at time point t is in turn parameterized as a mixture of K clusters weighted by cell proportions at the time point π_t as

$$W_{it} \sim \mathcal{TN}(\boldsymbol{\pi}_t \cdot \boldsymbol{\eta}_i, \sigma_W), \quad \boldsymbol{\pi}_t, \boldsymbol{\eta}_i \in \mathbb{R}^K \tag{8}$$

Contrary to the gene dosage, we expect transcription profiles of cells to change across time points leading to phenotypic plasticity. By default we set $\sigma_W = 0.05$ to account for the gene-specific variability in copy numbers. For gene i, cluster k, and time point t, we model the transcription rate of cells in cluster k as

$$c_{tik} \sim \Gamma(\alpha_{ti}, \frac{1}{\eta_{ik}}), \tag{9}$$

where the gene dosage η is the scale parameter of the Gamma distribution. The shape is given by a free parameter α_{ti}. As a result, we establish a hierarchical structure in which the gene dosage impacts both mean and variance of the transcriptional rate. Finally, to obtain the single-cell RNA counts X, we identify the cluster that each cell belongs to and the corresponding transcription rate vector. We also account for the effect of library size and cluster-specific bias in gene dosage. We sample X at time point t as

$$X_t \sim \mathrm{Poisson}(c_t[z_t] \odot \mathbf{l}_t \odot \mathbb{E}_g[\eta]) \tag{10}$$

$$c_t, z_t \in \mathbb{R}^{N\times G}, \quad \eta \in \mathbb{R}^{G\times K}, \quad \mathbf{l}_t \in \mathbb{R}^N \tag{11}$$

where z_t is the cluster membership matrix at time point t.

Variational inference

Variational family for lower Cholesky factors. We first construct the variational distribution of the lower Cholesky factors (L) of the covariance of gene dosage (η). Formally, we recover the lower Cholesky factor of a $K \times K$ matrix, denoted as L. We need $\binom{k}{2}$ free parameters which we denote as y. We assume the free parameters follow a diagonal normal distribution with prior $\mathbf{y} \sim \mathcal{N}^{\binom{k}{2}}(\mathbf{0}, I)$. We learn themeans and the variances of y as free parameters. From y, we can construct latent lower triangular matrix $\zeta \in \mathbb{R}^{K\times K}$ by filling the below diagonal entries by row with transformed y

$$\zeta = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ \tanh(y) & & 0 & \\ & & & 0 \end{pmatrix}$$

Then we can recover the cholesky factor L as

$$L_{ij} = \begin{cases} 0 & \text{if } i < j \\ \sqrt{1 - \sum_{j'<j} L_{ij'}^2} & \text{if } i = j \\ \zeta_{ij}\sqrt{1 - \sum_{j'<j} L_{ij'}^2} & \text{if } i > j \end{cases}. \tag{12}$$

Denote the function mapping from y to L as $g : \mathbb{R}^{\binom{k}{2}} \mapsto \mathbb{R}^{K\times K}$, and the variational distribution of L is given by

$$q(L; \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y) = g\left(q(\mathbf{y}; \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y)\right) = g\left(\mathcal{N}^{\binom{k}{2}}(\mathbf{y}; \boldsymbol{\mu}_y, \mathrm{diag}(\boldsymbol{\sigma}_y))\right) \tag{13}$$

General variational family structure. For notation simplicity we drop dependency on observed random variables. The variational family of Echidna is given by

$$q(\eta, c, L, \boldsymbol{\sigma}_\eta, \alpha) = q(\eta \mid \boldsymbol{\sigma}_\eta, L)\, q(c \mid \alpha, \eta)\, q(L; \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y)\, q(\boldsymbol{\sigma}_\eta) \tag{14}$$

which can be factorized into

$$q(\eta, c, L, \boldsymbol{\sigma}_\eta, \alpha) = q(\boldsymbol{\sigma}_\eta)\, q(L; \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y) \prod_{i=1}^{G} q(\boldsymbol{\eta}_i \mid L, \boldsymbol{\sigma}_\eta) \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{i=1}^{G} q(c_{tik} \mid \alpha_{ti}, \eta_{ik}) \tag{15}$$

Mean-field variational inference is applied to learn posterior for each latent random variable. We parameterize the variational distributions as

$$q(\sigma; a, b) = \prod_{k=1}^{K} \Gamma^{-1}(a_k, b_k) \tag{16}$$

$$\Sigma = LL^T, \Sigma \leftarrow \Sigma - I + \mathrm{diag}(\sigma) \tag{17}$$

$$q(\eta_i | L, \sigma) = \mathcal{N}^K(\mu_\eta, \Sigma) \tag{18}$$

$$q(c_{tik} | \alpha_{ti}, \eta_{ik}) = \Gamma\left(\alpha_{ti}, \frac{1}{\eta_{ik}}\right), \tag{19}$$

where $a, b \in \mathbb{R}_+^K$ are variational parameters for the inverse gamma distribution on σ. Notice that although the notations of variational parameters remain unchanged for clarity, these are different parameters compared to the parameters in the generative process.

The variational inference objective. Variational inference aims to minimize the KL divergence between the variational posterior and the true posterior, which is equivalent to maximizing evidence lower bound. We will denote the variational family as q_θ where θ is the set of all variational parameters and generative densities as p_ϕ where ϕ is the set of all generative parameters. The evidence lower bound is given by

$$\begin{aligned}\log p_\phi(X, W) \geq\ & \mathbb{E}_{q_\theta}[\log p_\phi(X|c, \eta) + \log p_\phi(W|\eta)] \\ & - D_{\mathrm{KL}}(q_\theta(c|\alpha, \eta) || p_\phi(c|\alpha, \eta)) - D_{\mathrm{KL}}(q_\theta(\boldsymbol{\sigma}_\eta) || p_\phi(\boldsymbol{\sigma}_\eta)) \\ & - D_{\mathrm{KL}}(q_\theta(L; \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y) || p_\phi(L)) - D_{\mathrm{KL}}(q_\theta(\eta|\boldsymbol{\sigma}_\eta, L) || p_\phi(\eta|\boldsymbol{\sigma}_\eta, L))\end{aligned}$$

Again, we note that parameters of q_θ and p_ϕ are different sets of parameters despite the identical notations.

GMM model to determine cluster-specific neutral CNA

To estimate neutral CNA within each cluster from the learned gene dosage(η), we fit a 1D GMM to account for the different levels of CNA signal. More specifically, for the kth cluster, the gene dosage vector is given by $\boldsymbol{\eta}_k = [\eta_k^1, \eta_k^2, \ldots, \eta_k^G]$. With P components, the probability density function of 1D GMMs are given by

$$p(\eta_{ik}) = \sum_{p=1}^{P} \lambda_p \mathcal{N}(\eta_{ik} \mid \mu'_p, \sigma'_p) \tag{20}$$

$$\sum_{p=1}^{P} \lambda_p = 1 \tag{21}$$

The model is fitted with an expectation-maximization algorithm and we use the output to predict which component each η_k^j can be sampled from. We then estimate the neutral CNA for cluster k, denoted as η_k^{neut}, as

$$N_p = \sum_{i=1}^{G} \mathbb{I}(p_i = p) \tag{22}$$

$$\hat{p} = \mathrm{argmax}_p N_p \tag{23}$$

$$\eta_k^{\mathrm{neut}} = \mu'_{\hat{p}}. \tag{24}$$

We do this for every cluster and construct vector $\boldsymbol{\eta}^{\mathrm{neut}} = [\eta_1^{\mathrm{neut}}, \eta_2^{\mathrm{neut}}, \ldots, \eta_K^{\mathrm{neut}}]$. We can then shift the values of η by $\boldsymbol{\eta}^{\mathrm{neut}}$ to center neutral CNV at 0 for all clusters.

HMM model for state prediction

To predict CNV states from the inferred gene dosage values (η), we first incorporate genomic positional information by smoothing the posterior estimate of η with a 1D Gaussian filter per cluster along the genome. This approach accounts for gene level noise and enhances sequential dependency of neighboring genes. We then fit GMMs to the smoothed copy number estimates for each cluster, and adjust η so that neutral CNA is centered around zero (see the 'GMM model to determine cluster-specific neutral CNA' section). We assume that the most frequent GMM component is the neutral state. We acknowledge that this assumption may not be valid in cases of multiploidy, in which case this model may require additional supervision. With the smoothed and adjusted η, we fit a HMM with Gaussian emission probability for each cluster-specific gene dosage vector. More specifically, consider smoothed and adjusted gene dosage η_k in cluster k, the joint likelihood of η_k and the sequence of hidden states $s = [s_1, s_2, \ldots, s_G]$ is given by

$$p(\boldsymbol{\eta}_k, \mathbf{s}) = p(s_1) \prod_{i=2}^{G} p(s_i \mid s_{i-1}) \prod_{i=1}^{G} p(\eta_{ik} \mid s_i) \tag{25}$$

We fit the model with the Viterbi algorithm to find the state sequence $s^* = \mathrm{argmax}_s p(\boldsymbol{\eta}_k, s)$ that maximizes the joint likelihood. This state sequence is the predicted CNV states along genes (or genomic bins).

Clone structure reconstruction

To reconstruct clone structure, we leverage the correlation structure across clusters to construct more robust hierarchical cluster trees. The correlation matrix can be obtained by normalizing the learned covariance parameter Σ. If the learned covariance matrix is noisy due to low sequencing depth or technical artifacts, we estimate the empirical correlation matrix from the smoothed learned gene dosage η to obtain a denoised version. More specifically, small, spurious spikes or dips in correlation values are averaged out while pronounced correlation signals remain unchanged. To obtain the clone structure, we perform hierarchical clustering with Ward linkage with the covariance matrix of gene dosage. More specifically, the distance matrix for hierarchical clustering D is given by

$$D = 1 - \rho \tag{26}$$

$$\rho = S^{-1} \ \Sigma \ S^{-1}, \tag{27}$$

where ρ is the correlation matrix and S is the diagonal matrix of standard deviations(square root of the diagonal elements of Σ). If we choose to estimate empirical correlation, the correlation matrix ρ is given by:

$$\hat{\Sigma} = \frac{1}{G-1} \eta^{\mathrm{T}} \eta \tag{28}$$

$$\rho = \hat{S}^{-1} \hat{\Sigma} \hat{S}^{-1} \tag{29}$$

where η is the learned gene dosage, $\hat{\Sigma}$ is the empirical covariance of η and $\hat{S}$ is the diagonal matrix of empirical standard deviations. We can then establish a threshold, either manually or based on the cophenetic distance of the hierarchical dendrogram, to derive flat clusters, which are subsequently identified as clones.

GDX

To understand the relationship between gene dosage (η) and transcriptional rate (c), we quantify statistical dependency between the random variables in the model. Such a relationship can be biologically interpreted as GDX, which reflects how much gene expression changes depending on genetic level alterations. Genes whose expression changes can be well correlated with genetic variations such as copy number (that is, high dosage effect), can be interpreted as intrinsic programs. Conversely, genes whose changes in expression do not depend on genetic variations (that is, low dosage effect) are likely to be driven by extrinsic or nongenetic effects. To quantify such a relationship, we

develop the following statistical framework. Consider the learned posterior distribution of c in one time point, we can factorize the densities to K different densities in each cluster k

$$q(c \mid \alpha, \eta) = \prod_{k=1}^{K} q_k(\mathbf{c}_k \mid \alpha, \boldsymbol{\eta}_k)$$

We then sample from the cluster-specific posterior distributions and get samples of η and c for all genes local to cell clusters. These samples live in $\mathbb{R}^G$ so they are per-gene measurements. We then compute variance shift from neutral copy number state as the GDX s_k in cluster k across these posterior samples to quantify signed normalized effect of a copy number change on expression for each gene for a specific phenotypic state k

$$s_k = \frac{\mathrm{Var}(q_k(\mathbf{c}_k \mid \tilde{\boldsymbol{\eta}}_k)) - \mathrm{Var}(q_k(\mathbf{c}_k \mid \boldsymbol{\eta}_k^0))}{\mathbb{E}[\mathrm{Var}(q_k(\mathbf{c}_k \mid \boldsymbol{\eta}_k))] + \mathrm{Var}(\mathbb{E}[q_k(\mathbf{c}_k \mid \boldsymbol{\eta}_k)])}$$

This metric allows us to score each gene in each subpopulation based on how its transcriptional variability is induced. We construct differential enrichment tests to capture the variation of plasticity-inducing mechanisms with gene signatures and connect these signatures to clinical outcomes such as clonal expansion or treatment response.

Permutation testing to identify significantly differential genes

Patient-wise cluster adjustment. To balance the number of clusters contributed by each patient, we resample exactly m clusters per patient where m is the median number of clusters across patients. Let $\mathcal{P}$ be the set of patient identifiers and, for each $p \in \mathcal{P}$,

$$I_p = \{\, i \mid \text{patient}\,(i) = p\},\qquad n_p = |I_p|.$$

The adjusted index set for patient p is

$$J_p = \begin{cases} \text{Sample}(I_p,\ m,\ \text{without replacement}), & n_p \geq m, \\ \text{Sample}(I_p,\ m, \text{with replacement}), & n_p < m. \end{cases}$$

The final dataset comprises $J = \bigcup_{p\in\mathcal{P}} J_p$ with total size $N = m\,|\mathcal{P}|$.

Feature matrix and binary labels. We obtain the numeric feature matrix $X \in \mathbb{R}^{N\times d}$. The categorical class label is mapped to a binary vector

$$y_i = \begin{cases} 0, & \text{if the sample is 'label 0'}, \\ 1, & \text{if the sample is 'label_1'}. \end{cases} \quad i = 1, \ldots, N.$$

Observed test statistic. For each feature $j = 1, \ldots, d$ we measure the difference in class- conditional means as

$$t_j^{\text{obs}} = \frac{1}{n_1} \sum_{\substack{i \in J \\ y_i = 1}} X_{i,j} - \frac{1}{n_0} \sum_{\substack{i \in J \\ y_i = 0}} X_{i,j},$$

where $n_1 = \sum_i 1(y_i = 1)$ and $n_0 = \sum_i 1(y_i = 0)$.

Permutation distribution. Let K be the number of label permutations. For each $k = 1, \ldots, K$ draw a random permutation π_k of $\{1, \ldots, N\}$ and define permuted labels $y_i^{(k)} = y_{\pi_k(i)}$. The corresponding test statistic is

$$t_j^{(k)} = \frac{1}{n_1} \sum_{\substack{i \in J \\ y_i^{(k)} = 1}} X_{i,j} - \frac{1}{n_0} \sum_{\substack{i \in J \\ y_i^{(k)} = 0}} X_{i,j}.$$

Two-sided permutation *P* value. The P value of our statistical analysis is given by

$$p_j = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}(|t_j^{(k)}| \geq |t_j^{\text{obs}}|), \qquad j = 1, \ldots, d.$$

Benjamini–Hochberg FDR correction. To control the false-discovery rate (FDR) at $\alpha = 0.05$, sort $p_{(1)} \leq \cdots \leq p_{(d)}$ and define thresholds $q_{(k)} = \alpha\, k/d$. Let $k^* = \max\{k : p_{(k)} \leq q_{(k)}\}$. All hypotheses with $p_{(k)} \leq q_{(k)}$ for $k \leq k^*$ are declared significant, and the corrected q-values are

$$\tilde{p}_{(k)} = \min_{l \geq k} \frac{d}{l}\, p_{(l)}, \qquad \tilde{p}_j = \tilde{p}_{(k)} \text{ where } j = j_{(k)}.$$

Gene signature enrichment based on GDX

To quantify gene signature enrichment and its statistical significance, we first rank all genes with the GDX score (for details see the 'GDX' section). Denoting the set of all ranked genes G and the gene signature $S \subset G$, we compute an enrichment score (ES) to reflect how genes in S are distributed among the ranked set G. The enrichment score is the absolute maximum deviation of the running-sum scores, namely

$$\mathrm{ES}(S) = \max_{1 \leq i \leq |G|} |R(i)|, \tag{30}$$

where $|G|$ is the number of genes in the ranked gene set and $R(i)$ is the running-sum score at position i, which is given by

$$R(i) = \sum_{j=1}^{i} \delta(g_j, S) \tag{31}$$

$$\delta(g_j, S) = \begin{cases} \frac{|\mathrm{GDX}(g_j)|}{\sum_{g \in S} |\mathrm{GDX}(g)|} & \text{if } g_j \in S \\ \frac{-1}{|G| - |S|} & \text{if } g_j \notin S \end{cases}. \tag{32}$$

To assess the statistical significance of enrichment score for gene signature S, we generate a null distribution by permuting the GDX scores and recomputing ES. The nominal P value is then given by

$$P\text{value} = \frac{\#(\mathrm{ES}_{\text{null}} \geq |\mathrm{ES}(S)|)}{\#\text{permutations}}. \tag{33}$$

We finally adjust the P values for multiple hypothesis testing with Benjamini–Hochberg correction.

Preprocessing of hybrid BAG-seq data

Paired snDNA/snRNA data along with corresponding WGS were obtained from ref. 40. The WGS data were mapped from genomic regions to genes to fit the standard Echidna expected inputs. Similarly, for ground-truth validation. the snDNA was mapped from the bins defined by ref. 40 into individual genes. We compared the result of this standard preprocessing against CHISEL to verify our decision in benchmarking (Supplementary Note D and Supplementary Fig. 4a–c). To run Echidna, only a few hyperparameters have to be adjusted. We increased σ_y to 0.1 during variational inference to account for the relatively low cell count. We also changed the initialization of the mean of the prior of μ_η to be the mean of the mapped WGS data.

To generate separate DNA-layer and RNA-layer BAM files from the hybrid BAG-seq single-nucleus libraries, we first classified each varietal tag (UMI) using the molecule-level 'template table' obtained from ref. 40. For each molecule, exon-overlap fraction, splice-junction counts, number of genes matched and transcript-overlap metrics were evaluated to identify RNA-derived templates versus DNA-derived templates using the same rules defined in the hybrid BAG-seq publication. We applied these criteria to each template table and produced two lists containing the corresponding (cell barcode, UMI and chromosome) combinations for RNA-layer and DNA-layer molecules.

Using these barcode–UMI–chromosome lists, we parsed the original coordinate-sorted hybrid BAM file and extracted all alignments matching each barcode–UMI pair, producing two independent BAM files representing reads assigned to the RNA layer or the DNA layer, respectively. Both output BAMs retained the original read structure and were sorted and indexed for downstream analysis. The RNA-layer BAM file was used for downstream benchmarking of tools requiring BAM inputs.

Benchmarking details

Running benchmarking baseline models. The two most commonly used models for cell-level CNA calling are InferCNV and Numbat. Therefore, we choose to evaluate Echidna's performance against these methods. InferCNV uses scRNA-seq data with reference cells that do not have CNV events to identify single-cell level CNAs[35,41]. Numbat takes advantage of genotype information from DNA data and infers CNAs in a haplotype-aware fashion[36]. In addition, we also benchmark with more recent methods such as CopyKAT[42] and XClone[43] to demonstrate the broader performance advantage of Echidna. Here we detail the specifics of how we ran the benchmarking baseline on each dataset.

Running InferCNV: when running InferCNV (v.1.23.0) on the tumor 2 and DEFND-Seq datasets, we lowered the `cutoff` from 1 to 0.01 due to the sparseness of our count matrix and used default values for the rest of the parameters: `cluster_by_groups`, denoise, and HMM were all kept as true. For the tumor 5 dataset, we lowered the InferCNV `cutoff` even further to 0.001 because of the increased sparseness of the tumor 5 count matrix compared to that of the other datasets. We used the T cells as our reference for tumor 2 and tumor 5; all nontumor cells were used as reference for the DEFND-Seq data.

Running Numbat: tor the tumor 2 run of Numbat (v.1.4.2), the genome was changed to 'hg19' and the number of cores was increased from 4 to 20 to speed up the run, but all other default parameters were left unchanged. For the DEFND-Seq data, the genome was set to 'hg38', but the other parameters were identical to those used for tumor 2. For all runs, any nontumor cells in the dataset were used as reference. Numbat could not be run on the tumor 5 dataset due to a data quality issue.

Running CopyKAT: when running CopyKAT (v.1.1.0), we used the same parameters for all datasets. We lowered `ngene.chr` from 5 to 1 to avoid excessively filtering cells, and all other parameters were kept at their default values.

Running XClone: before running XClone (v.0.3.9), we used the xcltk module to preprocess our data. XClone itself has three different modules: read depth ratio (RDR), B allele frequency (BAF) and combine (which uses the results from both RDR and BAF).

When running the RDR module for all three datasets, we changed `cell_anno_key` and `marker_group_anno_key` to 'cell_type', `plot_cell_anno_key` to 'Type', lowered `top_n_marker` to 5 to exclude less informative genes, and changed the `start_prob` array to [0.3, 0.4, 0.3] to decrease bias for the neutral state.

Similarly, in the BAF module, we made the same change to `cell_anno_key` and `plot_cell _anno_key`. We also set `filter_ref_ave` to None to reduce extra filtering, concentration at 35.5 to avoid over-smoothing, `concentration_lower` at 20 and `concentration_upper` at 100 (following example runs on XClone documentation), `remove_guide_XY` to False, `WMA_window_size` at 6, and `CNV_N_components` at 5. These parameters were applied to all three datasets.

Finally, in the Combine module for each dataset, we changed `cell_anno_key` and `marker_group _anno_key` to 'cell_type' and we set `merge_loss` to False. For all three modules, any parameters not mentioned were kept at their default values.

Benchmarking clone reconstruction. Echidna clones are called by performing Ward linkage based hierarchical clustering with the inferred covariance Σ_η (for details see the 'Clone structure reconstruction' section). Numbat clones are directly obtained from the output of the last iteration of phylogeny optimization. InferCNV and CopyKAT clones are identified by first pseudo-bulking the single-cell CNA estimates at the cluster level, followed by applying K-means clustering. We assign K as the ground-truth number of clones identified from scDNA data. XClone clones are obtained by linearly transforming the single-cell-level posterior probability matrix of CNV states to a continuous CNA estimate matrix and perform the same pseudo-bulk and K-means process as InferCNV and CopyKAT. More specifically, denote the posterior matrix returned by XClone as $P \in \mathbb{R}^{C\times G\times 3}$, where C is the number of cells, G is the number of genes and the third dimension is the probability of loss, neutral and gain. We construct a state vector as

$$\mathbf{c} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \qquad c_1 = -1,\ c_2 = 0,\ c_3 = 1. \tag{34}$$

The expected copy number of gene j in cell i is in turn given by

$$S_{i,j} = \sum_{s=1}^{3} P_{i,j,s}\, c_s. \tag{35}$$

We then perform a column-wise standardization to get a single-cell-level continuous estimate of copy number as

$$\mu_j = \frac{1}{C}\sum_{i=1}^{C} S_{i,j}, \tag{36}$$

$$\sigma_j = \sqrt{\frac{1}{C}\sum_{i=1}^{C}\left(S_{i,j} - \mu_j\right)^2}, \tag{37}$$

$$\tilde{S}_{i,j} = \frac{S_{i,j} - \mu_j}{\sigma_j}. \tag{38}$$

The matrix $\tilde{S} \in \mathbb{R}^{C\times G}$ is used to perform pseudo-bulking and K-means clustering similar to ref. 22. The performance metrics used are purity, adjusted mutual information and adjusted rand index. Assume we have N cells and K clones. The purity is given by

$$\text{Purity} = \frac{1}{N}\sum_{k=1}^{K} \max_j |C_k \cap C_j|, \tag{39}$$

where C_k are cells belonging to clone k and C_j are cells belonging to ground-truth clone j. This score is biased toward more clones, but measures the number of cells in the intersection of each clone k with ground-truth clone j. To account for the number of clones bias, we introduce two metrics adjusted for chance. Given two sets of clone labels U and V, the AMI is given by

$$\text{AMI} = \frac{\text{MI}(U,V) - \mathbb{E}[\text{MI}(U,V)]}{\text{avg}(H(U),H(V)) - \mathbb{E}[\text{MI}(U,V)]}, \tag{40}$$

where $H(.)$ is the entropy. The ARI is given by

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}, \tag{41}$$

where RI is the ratio between the number of agreeing cell pairs and the total number of cell pairs. The pairs are constructed over identified clones and the ground-truth clones.

Benchmarking CNA profile estimation. As scDNA CNA profiles are binned[40], we apply the binning introduced in ref. 40 to the estimated cluster-level CNA profiles from all methods. For Echidna, we directly used the learned gene dosage parameter η. For InferCNV and CopyKAT, we pseudo-bulk the single-cell CNA estimates at the cluster level, with

each cluster's CNA profile defined as the mean of the CNA profiles of all cells within the cluster. For Numbat, we expanded the clone-level pseudo-bulked total copy number ratio relative to diploid, where the estimated CNA profile for each cluster corresponds to the clone that is most prevalent within that cluster. For XClone, we use the continuous copy number estimate matrix derived from the returned posterior probability matrix and perform identical process to InferCNV and CopyKAT. For the derivation of the continuous copy number estimate matrix, see the 'Benchmarking clone reconstruction' section. All the CNA profiles are then compared to the scDNA CNA profile and reconstruction quality is measured by Pearson and Spearman correlation coefficient. For each method the CNA profile matrix is flattened into a vector $\mathbf{x} \in \mathbb{R}^{K \cdot B}$ where B is the number of genomic bins and K is the number of cell clusters. The CNA profile from scDNA data is similarly flattened into a vector $\mathbf{y} \in \mathbb{R}^{K \cdot B}$. The Pearson correlation coefficient is given by

$$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{42}$$

$$r = \text{sign}(m) \cdot \frac{\text{cov}(x, y)}{s_x s_y}, \tag{43}$$

where x_i and y_i are individual components of x and y, $\bar{x}\bar{y}$ are means of x and y, m is the linear regression slope and s_x, s_y are standard deviations of x and y.For Spearman correlation, we first obtain the rank vectors as

$$R_i = \text{rank}(x_i), \qquad S_i = \text{rank}(y_i), \qquad i = 1, \ldots, K \cdot B. \tag{44}$$

The Spearman correlation coefficient is given by

$$\rho_S = \frac{\sum_{i=1}^{K \cdot B} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{K \cdot B} (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^{K \cdot B} (S_i - \bar{S})^2}}. \tag{45}$$

Benchmarking CNV state prediction. We evaluate CNV state performance in each cluster by predicting the states (amplification or deletion) for all non-neutral bins of scDNA. We obtain the ground-truth CNV states by running HMMs on the scDNA CNA profile. Echidna states are predicted by fitting HMMs on gene dosage η as well (for HMM details see the 'HMM model for state prediction' section). InferCNV and CopyKAT states are determined by thresholding each cluster's CNA profile using the cluster-specific standard deviation. Specifically, a bin is classified as amplification if it is one standard deviation or more above the mean, and as deletion if it is one standard deviation or more below the mean. Numbat states are derived directly from the clone-level pseudo-bulk profile output. For each bin, the maximum a posteriori (MAP) CNV states of all SNPs within the bin are collected and majority voting is used to determine the bin's final CNV state. XClone CNV states are directly extracted from the CNV state posterior probability matrix. The prediction performance in each cluster is measured by standard metrics, namely precision, recall, true positive rate and false positive rate.

WGS data collection for glioblastoma

We obtained WGS data from a cryopreserved glioblastoma specimen from which joint snDNA/snRNA-seq data had previously been obtained with DEFND-seq[32]. Tissue was processed with a Dounce homogenizer and genomic DNA was extracted using a Quick DNA MiniPrep kit (Zymo D3024). We prepared a sequencing library from purified genomic DNA using the xGEN DNA Library Kit EZ (Integrated DNA Technologies, 10009864 and 10009914) with three cycles of PCR. Libraries were sequenced on the Element Aviti platform with 2 × 71 base-pair paired-end reads using a medium output kit (Element Biosciences, 860-00014) after loading the flow cell with the DNA library at 11 pM. We aligned the resultant sequencing data to the human reference genome (GRCh38) using the Burrows–Wheeler Aligner (bwa-mem, v0.7.17-r1188).

Application to DEFND-Seq glioblastoma

We applied Echidna to previously published matched snDNA/snRNA-seq data from ref. 32. WGS data were processed using ichorCNA (Extended Data Fig. 5a). We set `ploidy` to 2, `normal` to 0.1–0.2, `maxCN` to 50, `estimatePloidy` to False, `txnE` to 0.9999 and `txnStrength` to 10,000. Parameters were selected with the prior understanding that focal amplifications of MDM2 and EGFR should be preserved.

Subsequently, we ran Echidna on all tumor cells from the sample, setting a single time point and using inverse_gamma = True, patience = None, n_steps = 6000, lr = .1 and q_corr_init = 1e − 2. In this case, we elected to run the model without early stopping as this dataset was more difficult to fit. We instead evaluated model fit by ensuring the known genes from ref. 32 exhibited correct patterns (EGFR and PDGFRA) between clones (Fig. 4b,e). We then called clones using Echidna with threshold =0.025 (Fig. 4a), and shifted each cluster's inferred CNA using the GMM inferred neutrals, with three components (Fig. 4d).

Differential gene expression testing was performed between ground-truth clones using Scanpy, with a Wilcoxon test and default parameters (Fig. 4f)

Application to metastatic organotropism data in PDAC

We applied Echidna to data from ref. 44, featuring primary PDAC tumors with eventual metastasis to the liver and the lung niche. The data consisted of 17 patients, on which we received paired WGS and snRNA-seq. We mapped the given ichorCNA results representing the bulk genomic data to genes, and ran Echidna individually on each patient, with patience = −2 and 500 steps. The number of GMM components was tuned for each patient based on the underlying distribution of inferred CNAs. We performed permutation testing between liver and lung metastatic recurrence (MR) clusters, to identify genes that had a significant change in copy number between the two metastatic recurrent niches. This list was then compared to the liver and lung MR signatures presented in[44]. We confirmed that only a small number of genes in the signature had a significant change in CNA between the categories (Extended Data Fig. 6).

Application to NSCLC metastasis data

We applied Echidna to matched primary NSCLC and metastasized brain tumors described in ref. 45, consisting of 1,501 primary tumor cells and 1,581 brain metastasis cells. To prepare the data for use in Echidna, we normalized both samples together and performed Leiden clustering to get phenotypic states (Fig. 5a). Matched WGS data were processed by ichorCNA[45] and subsequently mapped to individual genes for use with Echidna. We ran Echidna with patient = None and 500 steps, and CNA calls were shifted per cluster using 4 GMM components per cluster. Here, we defined clones by computing a seaborn clustermap (Fig. 5b), using a correlation metric, on smoothed (sigma = 10) inferred CNAs, filtered to 5,646 genes using Echidna default HVG filtering. Clones were defined from the resulting dendrogram using a threshold of 0.08 (Fig. 5b).

DEGs in metastatic clones were computed using Scanpy, using a Wilcoxon test and default parameters (two sided). The density of these genes along the genome was inputted into the Find Peaks function of Scipy, with the prominence value set to 0.5. We defined a genomic hotspot to be a peak of DEGs over a region whose mean CNA among metastatic clones was greater than 0.5. The resulting list of DEGs present in genomic hotspots were further analyzed through gene set enrichment, which we performed through implementing the Fisher's exact test utilized by the DAVID functional annotation tool[69] to find enrichments in gene set enrichment analysis (GSEA) Hallmarks pathways. Specifically,

for each query of genes, we define LHas the number of genes from a known signature in the query list, LT as the length of the gene signature, PH as the number of genes from the signature in the full data and PT as the number of genes in the full data. We then apply Scipy's Fisher's exact test on the table [[LH,PH-LH],[LT-LH,PT-LT-(PH-LH)]] to obtain a statistic and *P* value describing how significantly the query list enriches for each signature of interest. The resulting values are visualized as a dot plot, with the Benjamini-Hochberg adjusted *P* value as the color and the percentage of genes found in the signature as the *x* axis (Fig. 5c).

Finally, we identified DEGs enriched in metastatic clones with statistically significant change in CNA between primary and metastatic clusters. To do so, we ran the above permutation test on inferred cluster-level CNAs, with 1,000 permutations. The resulting set of significantly different CNAs was overlapped with the set of metastatic DEGs, and relevant metastasis/invasiveness genes were selected from among them. We visualized the inferred *z*-scored CNAs for these genes (Fig. 5d) as well as the mean expression across all relevant metastasis genes (Fig. 5e).

Application to melanoma

We applied Echidna to a patient with melanoma from the KEYNOTE-001 clinical trial previously described in ref. 22. To prepare the data for use in Echidna, we normalized both treatment time points together and performed Leiden clustering to get phenotypic states. Matched WGS data were processed following the standard pipelines of Illumina and ichorCNA[22], and subsequently mapped to individual genes for use with Echidna. InferCNV and Numbat results were previously reported in ref. 22, which we include in Extended Data Fig. 8a–c as a comparison against Echidna.

We additionally ran XClone and CopyKat on this patient to compare with Echidna (Extended Data Fig. 9a–c). For XClone, we report the smoothed log ratio results from the RDR module, which was run using default configurations with the T and B cells from the dataset used as the reference. For CopyKAT, ngene.chr was decreased to 1 to avoid aggressive cell filtering, and all other parameters were left at default values. Each method was run separately for each of the two time points for which this patient was sampled. The results were subsequently concatenated and *K* means of $k = 4$ was applied to call clones across both time points together (following the process outlined in ref. 22).

For downstream analysis, we first computed DEGs by aggregating cells belonging to clones that Echidna classified as expanding or contracting. Specifically, a clone is labeled as expanding if the $\log_2$ fold change of the fraction of cells in that clone in the on time point divided by the fraction of cells in that clone in the pre time point is greater than 0.75, and contracting if the $\log_2$ fold change is less than −0.75 (if a clone was only present in the pre-therapy time point, it is automatically labeled as contracting, and vice versa if it is only present in the on time point). This resulted in 3,821 cells labeled as expanding and 9,577 labeled as contracting, with 4,854 genes resulting from filtering genes greater than the 75th quantile in variance. Differential expression testing was performed with Scanpy. Specifically, we selected the Wilcoxon rank-sum test (two tailed), with Benjamini–Hochberg *P* value correction.

For visualization in Fig. 6b and Extended Data Fig. 7f, we selected the top 100 significant DEGs in each condition, ranked by positive log fold change. The density of these genes along the genome was input into the Find Peaks function of Scipy, with the prominence value set to 0.4 for expanding and 0.3 for contracting. We defined a genomic hotspot to be a peak of DEGs over a region that contained a majority of amplified labels, as predicted by our HMM.

The significance of the GDX of all significant positive DEGs in each hotspot was compared against the full list of significant positive DEGs using the permutation test, described above. Within expanding clusters, we tested 194 hotspot DEGs against 1691 significant DEGs (filtered by corrected *P* value <0.05 and log fold change >0). Within contracting clusters, we tested 304 hotspot DEGs against 2,481 significant DEGs. The same permutation test was applied to study the significance of individual gene signatures in Supplementary Fig. 1. The genes in each tested signature may be found compiled in ref. 65, with the addition of our ICR signature[22], and were tested against a set of 2,913 genes obtained by filtering for genes greater than the 85th quantile of variance.

To compare the difference in GDX and expression in our previously published set of ICR genes[22], we constructed two sets of genes as follows: first, we limit the set of differentially expressed ICR genes (enriched in expanding clusters) to those with corrected *P* value <0.05 and log fold change >0. Then, we split the genes between ICR genes falling within a hotspot region and those not in a hotspot region. We next tested the GDX as well as gene expression of these two sets of genes in expanding and contracting clusters (we note that we only test ICR genes that were differentially expressed in expanding clusters because we did not expect these genes to play a notable role in contracting clusters). Specifically, to test the GDX, we constructed four lists of GDX values: GDX values for differentially expressed hotspot ICR genes in expanding clusters (23 genes across 4 clusters), GDX values for differentially expressed non-hotspot ICR genes in expanding clusters (117 genes across 4 clusters), GDX values for differentially expressed hotspot genes in contracting clusters (23 genes across 8 clusters) and GDX values for differentially expressed non-hotspot genes in contracting clusters (117 genes across 8 clusters). We report *P* values between groups of interest as shown in Fig. 6d using a two-tailed Mann–Whitney *U* test, with no *P* value correction due to the small number of tests.

Analysis of melanoma Spatial SlideSeq data

Spatial SlideSeq data previously published in ref. 22 was loaded into Squidpy[66] and normalized using Scanpy. To determine the spatial localization of clones, we identified the top 20 DEGs ranked by log fold change in clone 4 in the on-treatment time point. We colored the spatial plot by the *z* score of the average expression across these genes and visually identified the region of the plot that was enriched for the signature. We then computed the Moran's *I* score for spatial autocorrelation as implemented by ref. 66, on the set of highly varying genes returned by Echidna from the scRNA-seq. Genes from among those with the highest Moran's *I* score were manually selected based on biological interest and range of GDX scores, and their *z*-scored expression was visualized in the spatial context (Fig. 6e).

Implementation and usage

The model is implemented with the PyTorch based statistical programming framework Pyro. The package follows the schematic of related tools for single-cell analyses, namely the `scanpy` package[70] and `decipher`[37]. The package manages the user's analysis workflow through an `AnnData` object, systematically storing intermediate data and metadata as functions from the `.tl` submodule are called. Plotting functions in the `.pl` submodule automatically retrieve stored data from the `AnnData` object's `.uns`, `.var` and `.obs` attributes, simplifying data management for the user.

Typical usage begins by importing the package with `import echidna as ec`. The user has the option to toggle between single time point mode(if samples are collected at the same time point) and multitime point mode(if samples are collected at different time points). To ensure a balanced representation of all time points, we downsample cells from each time point to the minimum number of cells across time points for multitime point mode. The user trains the model by invoking `ec.tl.echidna_train(adata, w_df)`, which requires an `AnnData` object `adata` containing scRNA data and a pandas DataFrame `w_df` with copy numbers for each time point, indexed by gene names. With a trained model, the user calculates learned clones by calling `ec.tl.echidna_clones(adata)` and can then estimate growth status with `ec.tl.echidna_status`.

For investigating copy number aberrations across the genome, `ec.tl.echi_cnv(adata, genome)` provides a comprehensive solution. This function accepts a pandas DataFrame `genome` with columns 'chrom', 'txStart', 'txEnd' and 'geneName', and assigns amplification, neutral or deletion status to each gene within a learned clone. While a default genome is provided, users are encouraged to use a genome most appropriate for their specific dataset. The results can be visualized using `ec.pl.plot_cnv(adata)`. The package enables the calculation and visualization of GDXs through `ec.tl.gene_dosage_effect(adata)` and `ec.pl.plot_gene_dosage(adata)`, respectively. For a comprehensive understanding of the package's capabilities, users are directed to the example notebooks, which offer detailed, step-by-step walkthroughs.

Reporting summary

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

Data availability

The endometrial cancer hybrid data are from ref. 40 and available at NCBI Sequencing Read Archive (SRA) with accession code PRJNA773107(https://www.ncbi.nlm.nih.gov/bioproject/PRJNA773107). Matched WGS for the endometrial cancer patients has been deposited at NCBI SRA under accession number PRJNA1360783 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1360783; https://www.ncbi.nlm.nih.gov/sra?term=SRP644973). The melanoma data is published in ref. 22, with processed data deposited in the Gene Expression Omnibus, accession number GSE192402 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE192402). Raw data are available on dbGAP, accession number phs003097.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003097.v1.p1). The PDAC data is available from ref. 44, with raw and processed data available via GEO with accession numbers GSE253429 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE253429) (snRNA-seq) and GSE253430 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE253430) (lpWGS). The brain metastases data are published at ref. 45. Processed data from this study were deposited at NCB GEO, with accession number GSE223503 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE223503), and raw data uploaded to dbGAP, with accession number phs003865.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003865.v1.p1). The DEFND-seq data are from ref. 32 and in the in the Gene Expression Omnibus (GSE224149) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE224149). Source data are provided with this paper.

Code availability

The Python implementation of Echidna is available via GitHub at https://github.com/azizilab/echidna. The software version used for the results in this work is available via Zenodo at https://zenodo.org/records/15595657 (ref. 71). The notebooks to reproduce figures are available via GitHub at https://github.com/azizilab/echidna_reproducibility

References

69. Dennis, G. Jr et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, P3 (2003).
70. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
71. O'Brien, W. et al. azizilab/echidna: Echidna. *Zenodo* https://zenodo.org/records/15595657 (2026).

Acknowledgements

We thank M. Wigler for helpful data sharing, support and discussion. We also thank A. Califano, Y. Jin, M. Chalabi and A. Degroote for helpful discussions and feedback.

Author contributions

E.A. and B.I. conceived the study and provided overall supervision of the study. J.L.F, M.Z., W.O.B., J.D.M., N.B.-V., A.P., I.A. and E.A. designed and developed the Echidna framework. J.L.F, M.Z., W.O.B., J.D.M. and E.A. contributed to implementation. P.A.S., T.O. and S.L. collected data and aided in analyses. J.L.F., M.Z., W.O.B., J.D.M., S.C. and E.A. analyzed the data. J.L.F., M.Z., J.D.M., J.C.M., J.B., E.D.S., S.T., K.H.-H., B.I. and E.A. interpreted the data. J.L.F, M.Z., K.H.H., B.I. and E.A. wrote the paper. All authors reviewed, contributed to and approved the paper.

Funding

J.L.F. acknowledges support from the Columbia University Van C. Mow fellowship and the Columbia Avanessians fellowship. K.H.-H. was supported by the National Science Foundation Graduate Research Fellowship under grant number DGE-2036197. B.I. is supported by National Institutes of Health (NIH) grants, R37CA258829, R01CA280414, R01CA312637, U54CA274506, the Pershing Square Sohn Cancer Research Alliance Award, the Burroughs Wellcome Fund Career Award for Medical Scientists; a Tara Miller Melanoma Research Alliance Young Investigator Award; a Tara Miller Melanoma Research Alliance Team Science Award; the Louis V. Gerstner, Jr. Scholars Program; and the V Foundation Translational All Star Award. B.I. is a CRI Lloyd J. Old STAR (CRI5579). B.I. is supported by the Herbert Irving Comprehensive Cancer Center Human Tissue Immunology and Immunotherapy Initiative. B.I. and E.A. acknowledge support from NIH NCI grant R01CA266446, Columbia Research Initiatives in Science & Engineering (RISE) Award, and a pilot award from Herbert Irving Comprehensive Cancer Center (HICCC) NIH NCI grant U54CA209997. E.A. is supported by the NIH NCI grant R00CA230195, NIH NHGRI grant R01HG012875, R21HG012639, and grant number 2022-253560 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. S.L. acknowledges support by a grant from the Simons Foundation, Life Sciences Founders Directed Giving-Research (award number 519054 to M. Wigler); The Breast Cancer Research Foundation (BCRF, to M. Wigler) and by support from the Cold Spring Harbor Laboratory and Northwell Health Affiliation on Endometrial Cancers (awarded to M. Wigler).

Competing interests

B.I. is a consultant for or received honoraria from Volastra Therapeutics, Johnson & Johnson/Janssen, Novartis, GSK, Eisai, AstraZeneca and Merck, and has received research funding to Columbia University from Agenus, Alkermes, Arcus Biosciences, Checkmate Pharmaceuticals, Compugen, Immunocore, Regeneron, and Synthekine. B.I. is the scientific founder of Basima Therapeutics. PAS receives patent royalties from Guardant Health. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03279-4.

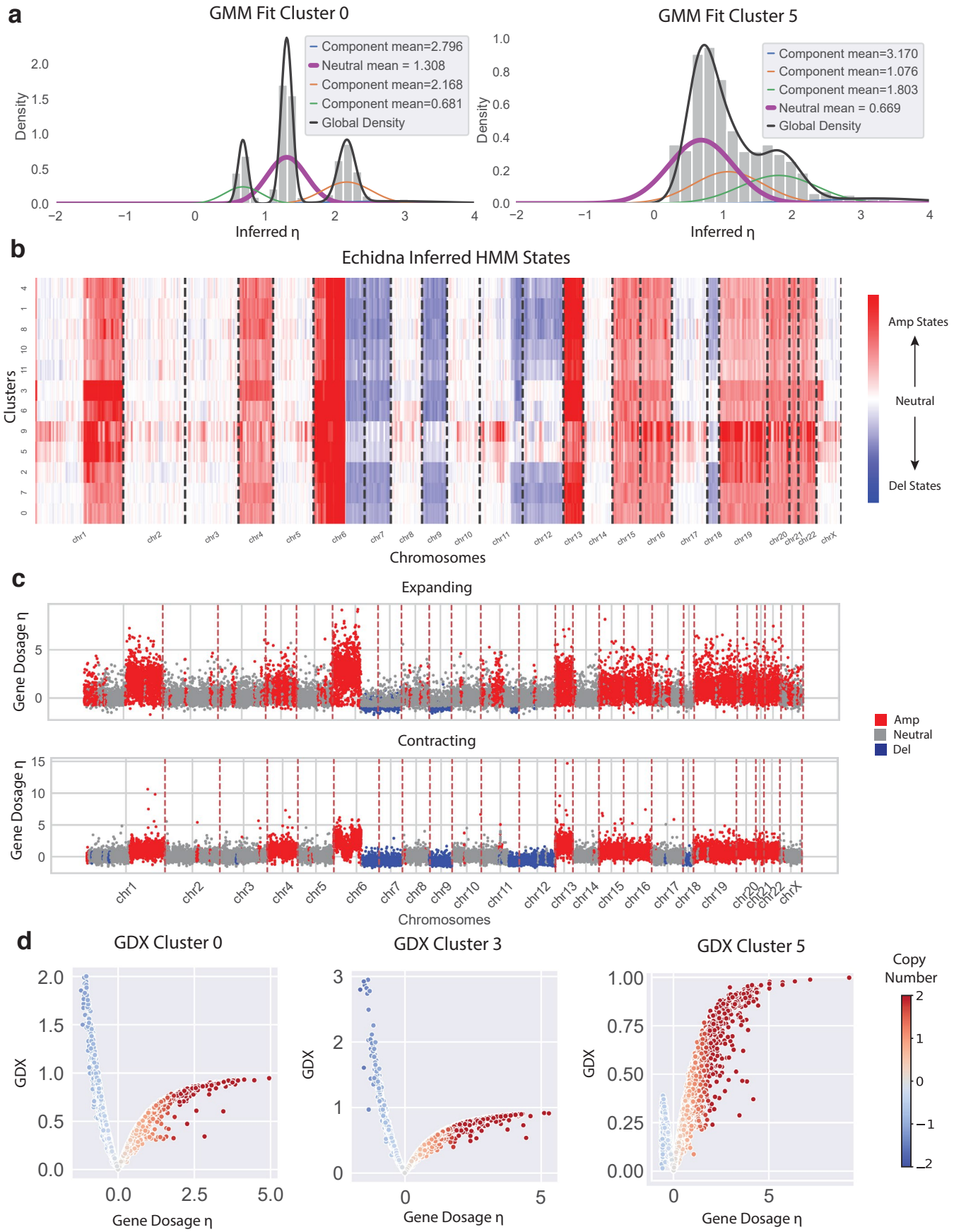
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03279-4.

Correspondence and requests for materials should be addressed to Benjamin Izar or Elham Azizi.

Peer review information *Nature Methods* Nature Methods thanks Kadir Akdemir, Woong-Yang Park and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-026-03279-4

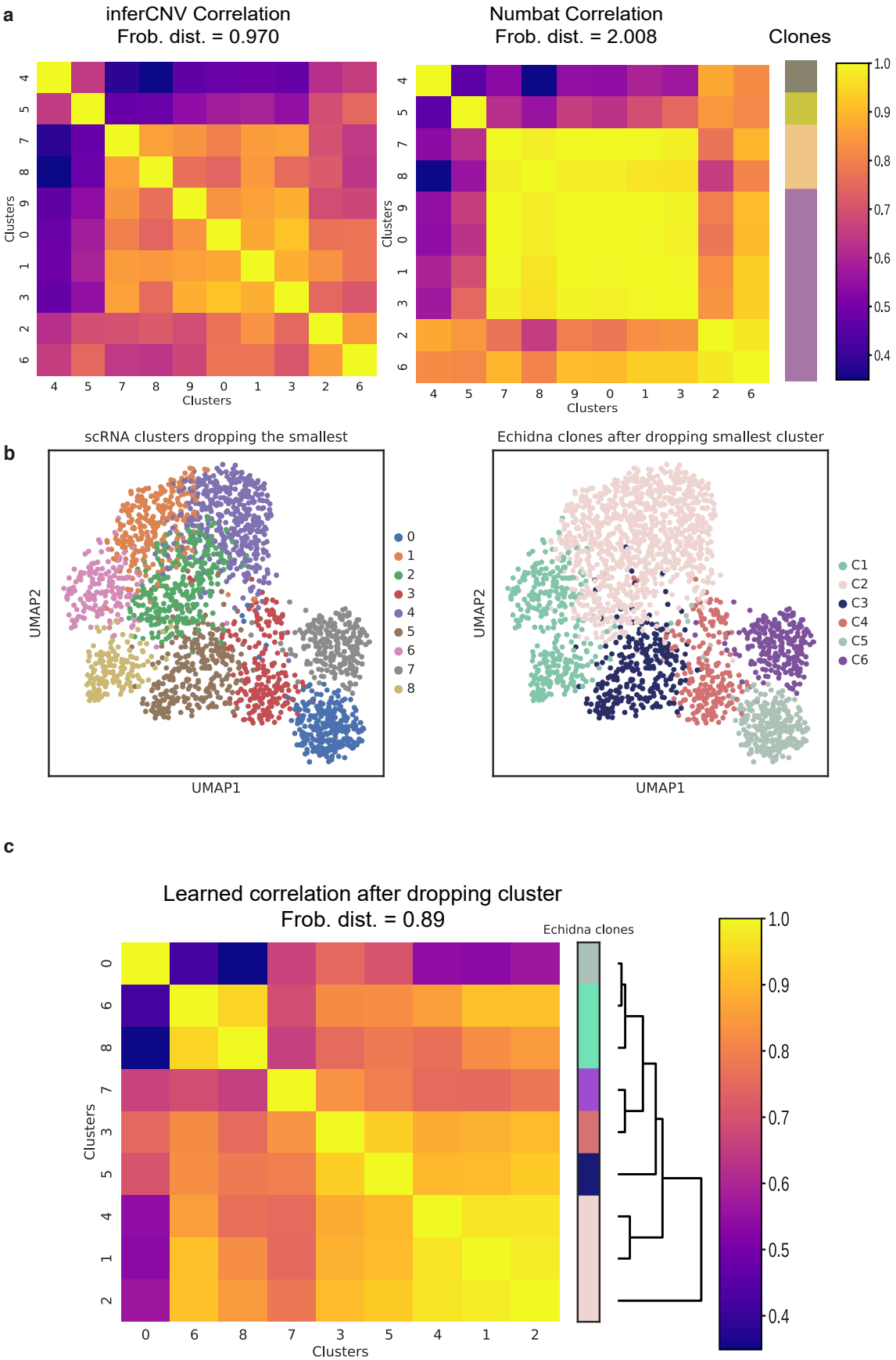


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-026-03279-4

Extended Data Fig. 1 | Echidna utilizes a Gaussian Mixture Model (GMM) and Gaussian HMM in post-processing steps. (**a**) Echidna fits a GMM to inferred copy number values per cluster. The GMM component with the most genes assigned (purple line) is used as the neutral mean. (**b**) HMM-inferred copy number states inferred from smoothed copy number values per cluster. Color represents different levels of amplified vs deleted states. (**c**) HMM-inferred copy number states based on learned gene dosage values (y- axis) in clusters grouped by temporal dynamics (expanding or contracting over time). Amp = amplification. Del = deletion. (**d**) Gene Dosage Effect (GDX) computed per cluster for a melanoma specimen as example. The scatterplot demonstrates the relationship between GDX (y-axis) and gene dosage (x-axis and color). Each dot represents a gene and is colored by gene dosage.

https://doi.org/10.1038/s41592-026-03279-4

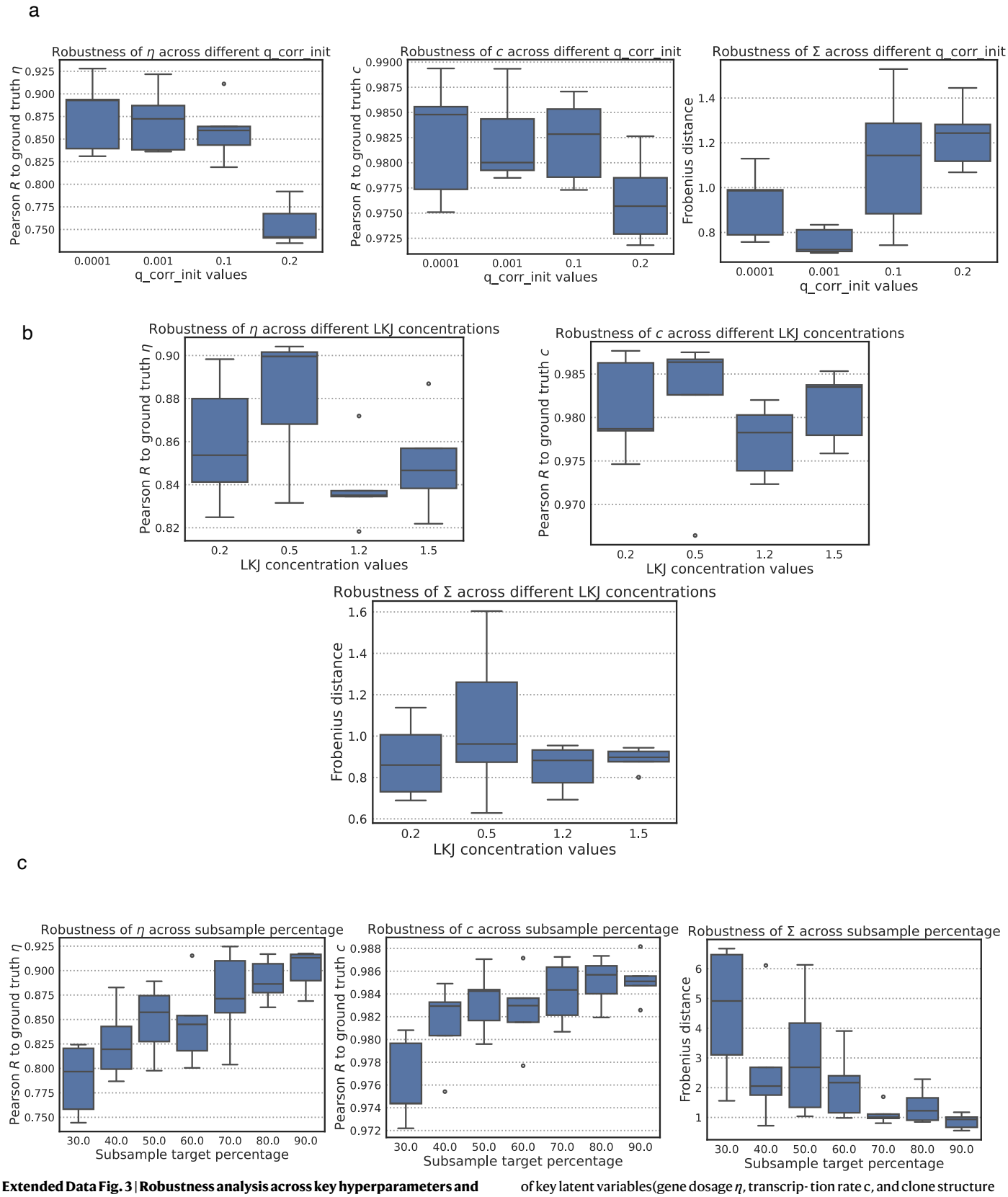


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41592-026-03279-4

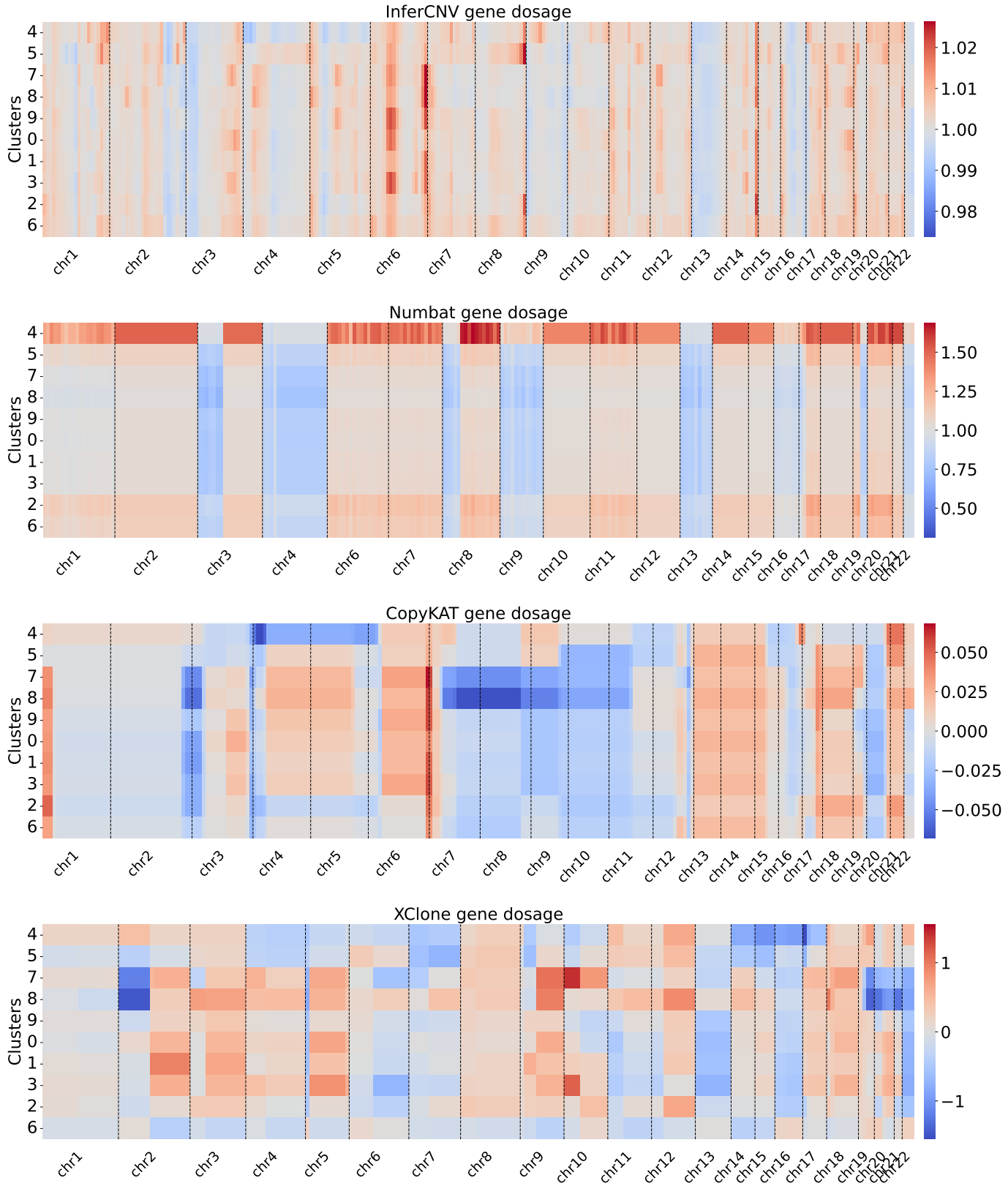
Extended Data Fig. 2 | Echidna's ability to reconstruct clones demonstrates robustness to noisy or small clusters. (**a**) Correlation between clusters recovered by InferCNV and Numbat for endometrial cancer tumor 2. Frobenius distances are computed to the ground truth correlation matrix from scDNA data. (**b**) UMAP of scRNA-seq data of tumor 2 while dropping the smallest cluster to test method robustness. Cells are colored by Leiden cluster labels (left) and clones recovered by Echidna (right). (**c**) Correlation matrix in- ferred by Echidna on tumor 2 data with the smallest cluster dropped. Echidna clone calls are shown. Dendrogram shows the grouping of cell clusters.

https://doi.org/10.1038/s41592-026-03279-4



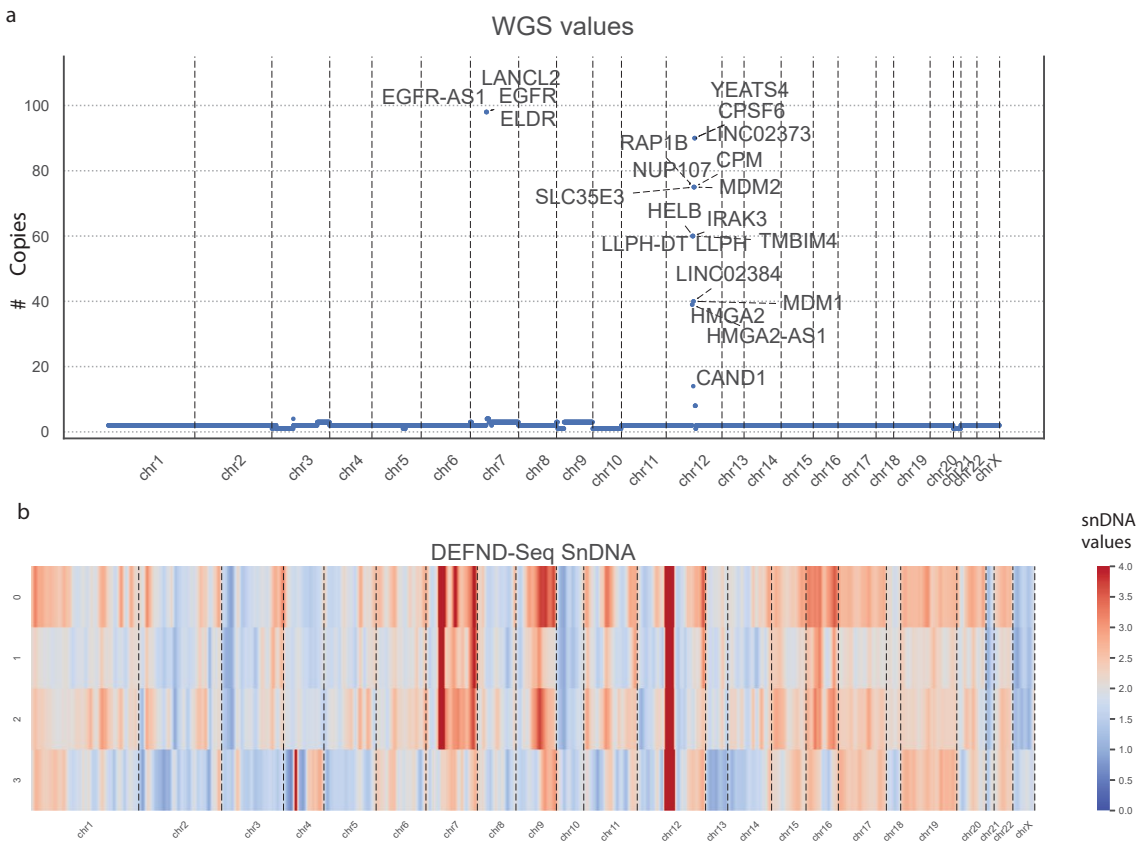
Extended Data Fig. 3 | Robustness analysis across key hyperparameters and low sample sizes. For box plots n = 20. The center line is median. The range of the box is the 25th and 75th percentile. The whiskers are 1.5 times IQR. (**a**) Robustness of key latent variables (gene dosage η, transcription rate c, and clone structure Σ) across different values of initial posterior correlation (q corr init). (**b**) Robustness of key latent variables(gene dosage η, transcrip- tion rate c, and clone structure Σ) across different values of LKJ concentrations. (**c**) Robustness of key latent variables (gene dosage η, transcription rate c, and clone structure Σ) across various low sample size scenarios. Subsample target percentage representing the percentage of cells remaining after down sampling.

https://doi.org/10.1038/s41592-026-03279-4



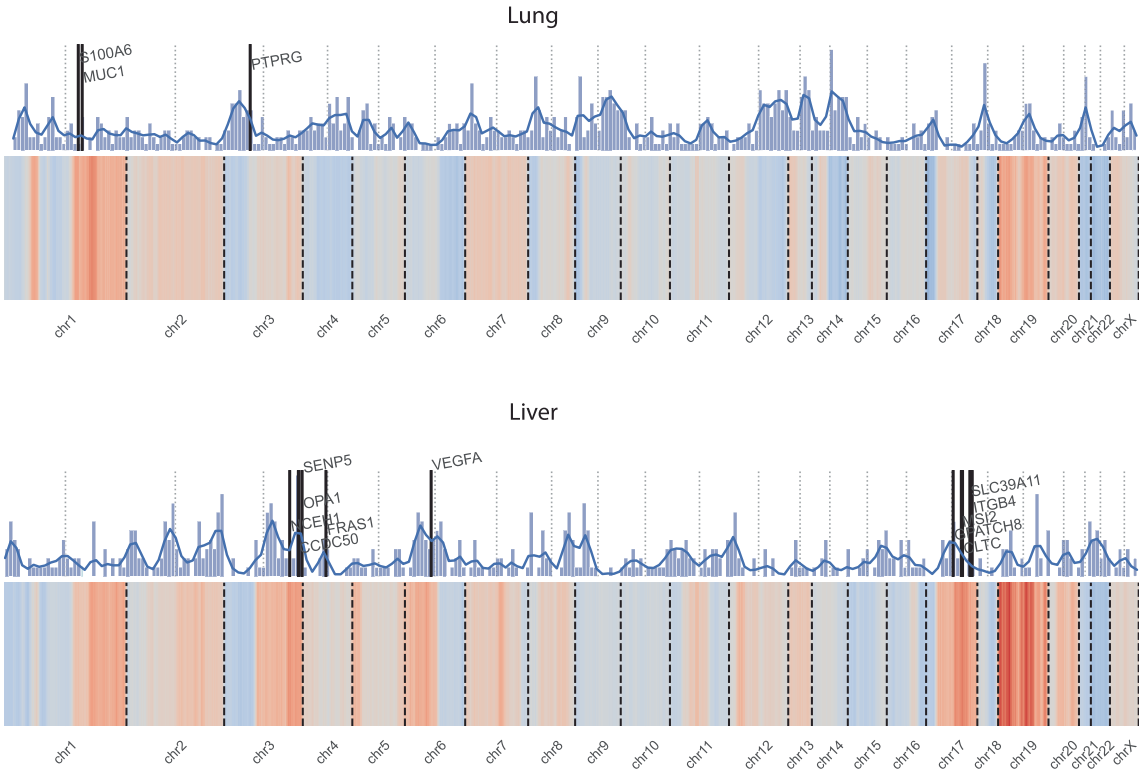
Extended Data Fig. 4 | Gene dosage profile estimation for endometrial cancer sample from all baseline models. (**a**): Cluster-level CNA profile inferred by InferCNV and Numbat for tumor 2 from[40]). Missing bins are due to Numbat's gene filtering.(**b**): Cluster-level CNA profile inferred by CopyKAT (**c**): Cluster-level CNA profile inferred by XClone.

https://doi.org/10.1038/s41592-026-03279-4



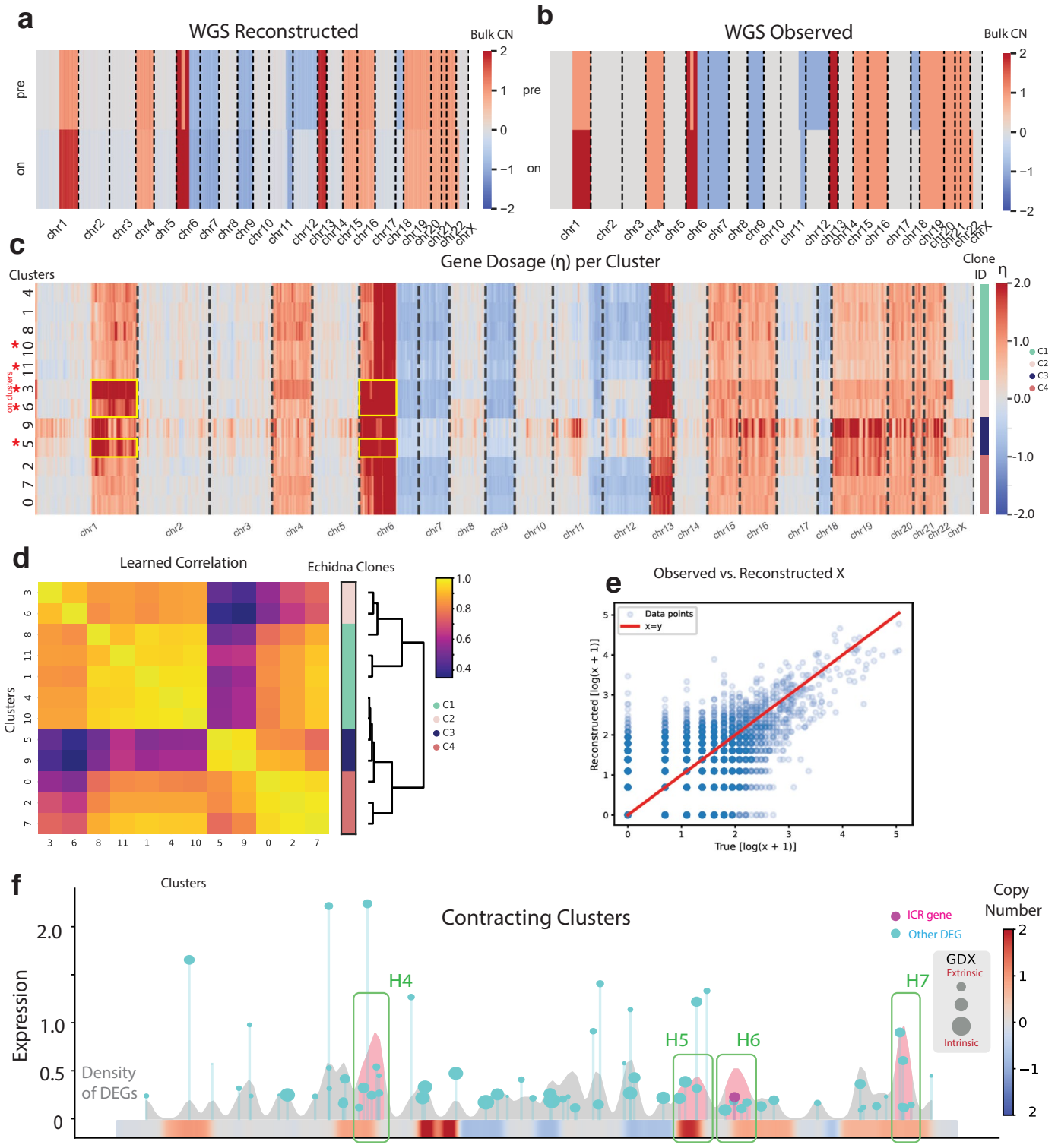
Extended Data Fig. 5 | DEFND-Seq dataset and matched WGS on glioblastoma. (**a**) WGS values per gene from the dataset of [32] as aligned by ichorCNA, ordered by genomic position. Genes with detected copies > 10 are labeled, revealing focal amplification events impacting key genes such as EGFR and MDM2. (**b**)snDNA values from DEFND-seq aligned to individual genes, showing impact of chr12 focal amplification across all tumor clusters.

https://doi.org/10.1038/s41592-026-03279-4



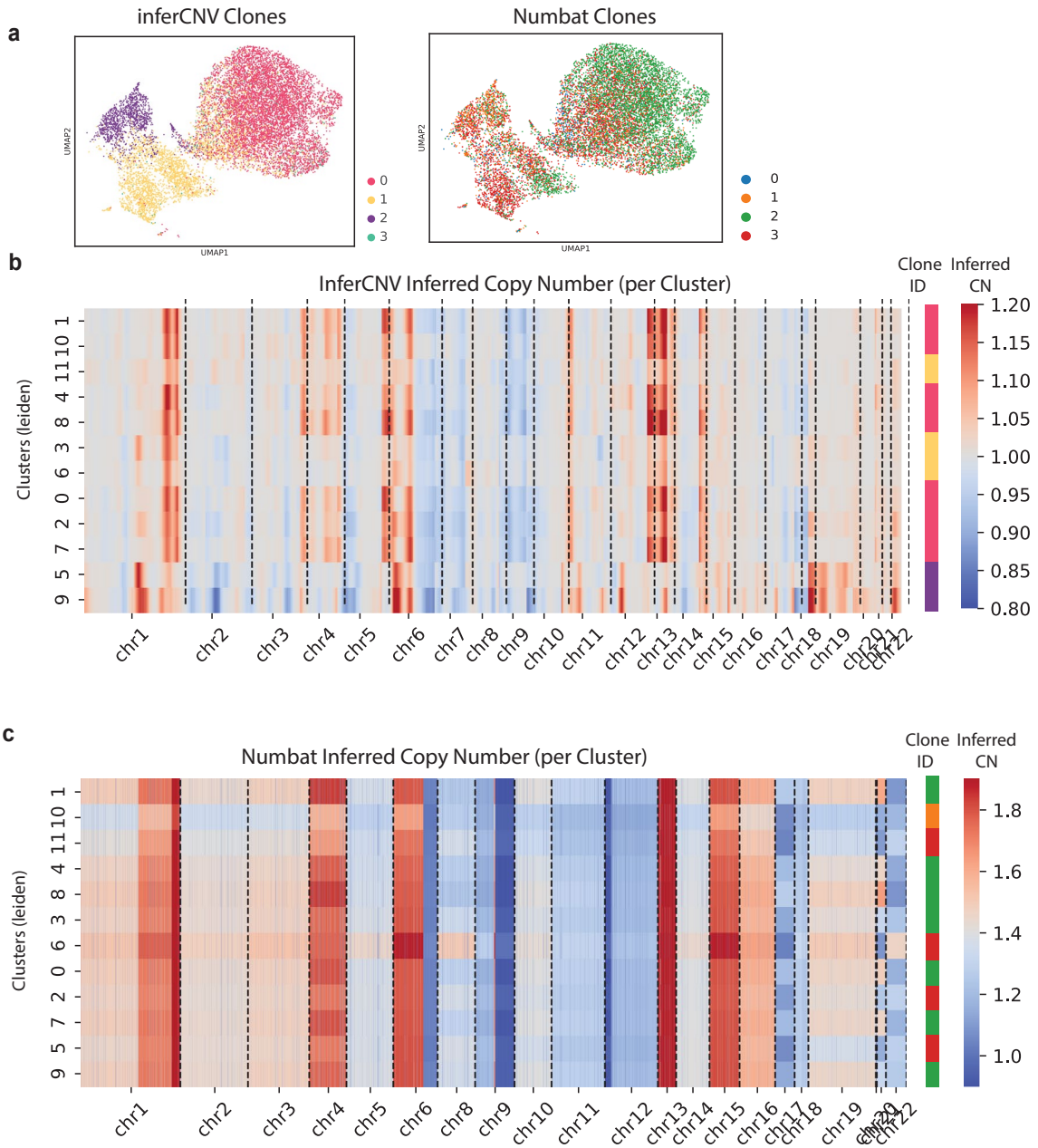
Extended Data Fig. 6 | Echidna validates extrinsic mechanisms in metastatic organotropism. Averaged inferred CNAs for patients with lung MR (top) and liver MR (bottom). Location of genes belonging to signatures defined in [44] for a density plot on top of each heatmap. Genes belonging to each signature with statistically significant CNA difference between categories are labeled.

https://doi.org/10.1038/s41592-026-03279-4



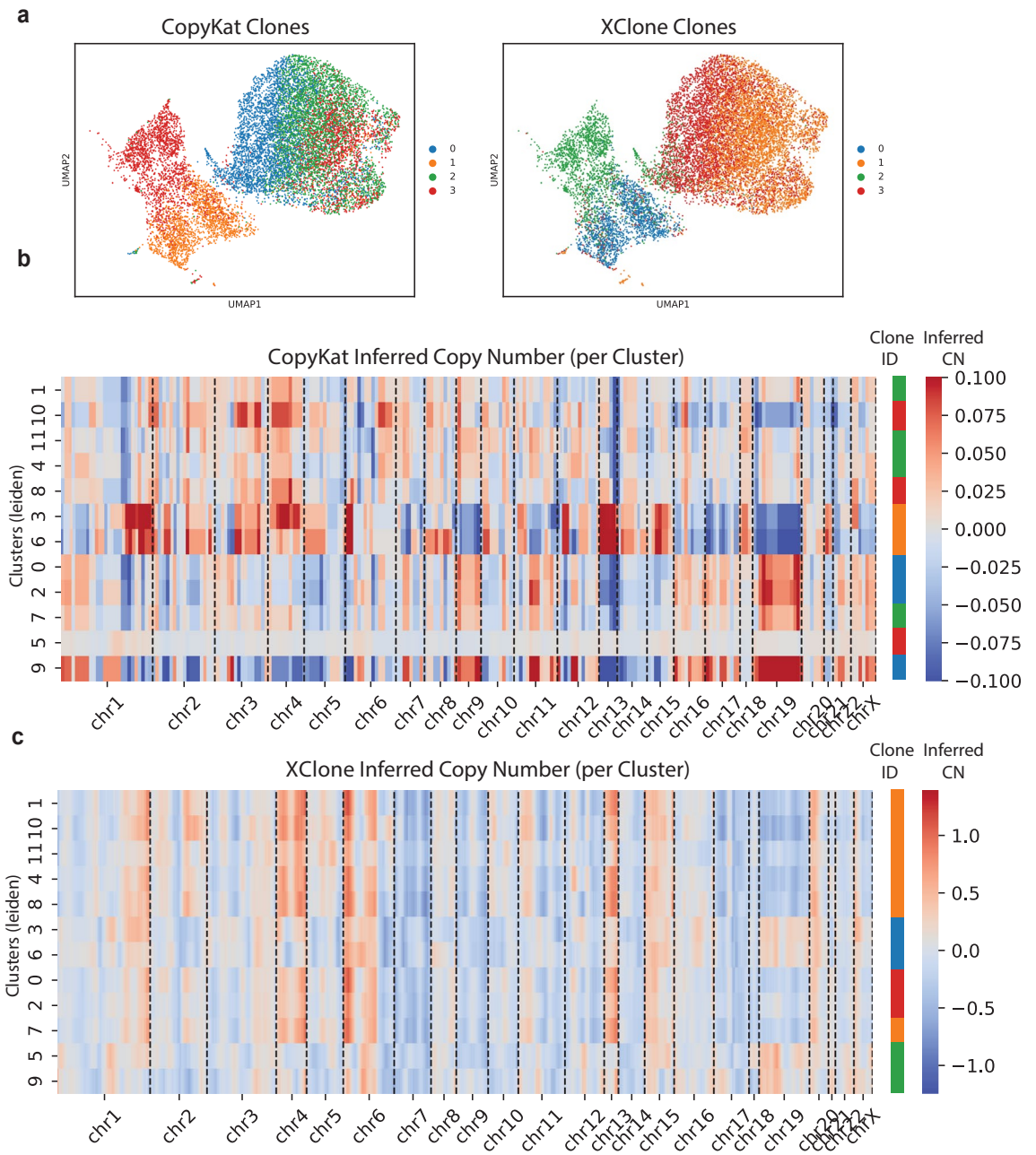
Extended Data Fig. 7 | Echidna applied to a multitime point dataset from a melanoma patient. Reconstructed (**a**) and observed (**b**) WGS data for each treatment time point in the melanoma patient. (**c**) Echidna-inferred gene dosage values per cluster. Red stars denote clusters in the on-treatment time point, while the right bar groups clusters into clones. Copy number states are shown in Extended Data Fig. 1b. (**d**) Inferred correlation matrix from Echidna, with clusters grouped by clonal structure. (**e**) Observed vs recon- structed snRNA-seq data (X). (**f**) Top differentially expressed (DE) genes in clusters contracting with treatment, visualized as lollipops with x-axis corre- sponding to genomic position. Density plot reflects the distribution of DE genes by genomic position. Height of each lollipop corresponds to the expression of each gene, while dot size reflects the absolute value of the GDX, such that larger dots denote stronger explained variance between gene dosage and expres- sion. Mean gene dosage is plotted underneath as a heatmap. ICR genes are highlighted in pink.

https://doi.org/10.1038/s41592-026-03279-4



Extended Data Fig. 8 | Performance of other methods on melanoma data (InferCNV and Numbat). (**a**) UMAP projection of scRNA-seq of tu- mor cells combined across two treatment timepoints similar to Fig. 6a with cells colored by clones defined from inferCNV results (left) and Numbat clones (right). (**b**) InferCNV-inferred copy number averaged across cells for each phe- notypic cluster. (**c**) Numbat-inferred MAP ratio to diploid averaged across cells for each phenotypic cluster.

https://doi.org/10.1038/s41592-026-03279-4



Extended Data Fig. 9 | Performance of other methods on melanoma data (CopyKat and XClone). (**a**) UMAP projection of scRNA-seq of tu- mor cells combined across two treatment timepoints similar to Fig. 6a with cells colored by clones defined from CopyKAT (left) and XClone (right). (**b**) CopyKAT-inferred copy number averaged across cells for each phenotypic clus- ter. (**c**) XClone-inferred RDR log ratios averaged across cells for each pheno- typic cluster.

nature portfolio

Corresponding author(s): Elham Azizi and Benjamin Izar

Last updated by author(s): 2026-04-26

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection requirements: bwa-mem=0.7.17-r1188
Data analysis	Version requirements for analysis and modeling: anndata=0.10.9, h5py=3.11.0, jupyter=1.0.0, leidenalg=0.10.2, matplotlib=3.10.8, numpy=1.26.4, pandas=2.2.2, python=3.11.9, scikit-learn=1.4.2, scipy=1.12.0, seaborn0.13.2, squidpy=1.6.1, scanpy=1.10.1, CopyKat=1.1.0, XClone=0.3.9, InferCNV=1.23.0, Numbat=1.4.2, hmmlearn=0.3.3, pyro-ppl=1.9.1, graphviz=0.21, requests=2.32.3, R=4.3.3 The Echidna package and requirements is available at https://github.com/azizilab/echidna.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The following datasets are all publicly avaialble: The endometrial cancer hybrid data is available at NCBI Sequencing Read Archive (SRA) with accession code (PRJNA773107). Matched WGS for the endometrial cancer patients has been deposited at NCBI Sequencing Read Archive (SRA) under accession number PRJNA1360783. The melanoma data has processed data deposited in the Gene Expression Omnibus, accession no. GSE192402. Raw data are available on dbGAP, accession no. phs003097.v1.p1. The PDAC data is available via GEO with accession numbers GSE253429 (snRNA-seq) and GSE253430 (lpWGS). The brain metastases data is deposited at NCB GEO, with accession number GSE223503, and raw data uploaded to dbGAP, with accession number phs003865.v1.p1. The DEFND-seq data is avaiable in the Gene Expression Omnibus (GSE224149).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No sex and gender information were required.
Population characteristics	All data analyzed in this study were obtained from previously published datasets, including single-cell and genomic profiles of cancer patient samples across multiple cohorts. Population characteristics, inclusion criteria, and recruitment procedures were defined by the original studies and are described in the corresponding publications cited in the manuscript.
Recruitment	Population characteristics, inclusion criteria, and recruitment procedures were defined by the original studies and are described in the corresponding publications cited in the manuscript.
Ethics oversight	GBM patient samples from which WGS was collected were procured from de-identified patients who provided written informed consent through a protocol approved by the Columbia University Institutional Review Board (IRB-AAAJ6163)

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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No primary data. Sample size recommendations follow the original published studies.
Data exclusions	Case studies were selected from original publications based on if a strong clonal substructure would benefit from study using Echidna
Replication	Data and analysis showed the reproducibility of findings. All other study design details are public and may be found in the original works.
Randomization	No randomization was performed in our computational experiments. We refer to the original study designs, which have all been published.
Blinding	No blinding was performed in our computational experiments, as the model is run on each patient separately. Covariates are not utilized in the training of the model.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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